# CLOSING THE LOOP ON HUMAN BEHAVIOR PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLM digital twins are usually scored against how closely people reproduce their own answers on a retest, as if retest agreement were a ceiling on individual accuracy. It is a floor, the accuracy of copying the earlier answer and hence a lower bound on the best accuracy attainable from information that includes that answer, and under stable propensities the nearest-neighbor inequality of Cover and Hart bounds that accuracy from above. We score LLM and non-LLM predictors of individual decisions on identical rows in the closed loop, where a predictor sees the person's earlier behavior and observes each decision after predicting it, on the Twin-2K-500 retest wave, three waves of the General Social Survey, Psych-101 and held-out Psych-201 experiments. Most of the individual signal our predictors extract lies in the person's most recent relevant behavior, and they differ in whether they anticipate a departure from it. On Twin-2K-500, dividing by retest agreement credits the item mode with 92.6% under the metric of the dataset's authors, every released twin falls below the item mode, and an LLM given the earlier answer matches copying only by repeating it. On cognitive tasks, conditioning on the previous choice is the largest gain for statistical predictors, which then predict repeats, and a recurrent network trained on other persons' choices and outcomes reaches the accuracy of the best LLM on Psych-101. Reading the person's raw transcript, gemini-3.6-flash detects more switches than CAOP-sel and the transition predictors at the same false-alarm rate, but on held-out Psych-201 experiments it is right on switches only slightly more often than the item mode. Predicting when and how a person departs from recent behavior remains open.

## 1 INTRODUCTION

Large language models conditioned on a person's survey answers or interview are increasingly used as digital twins that predict what that person will choose (Park et al., 2026; Toubia et al., 2025; Karetnikov et al., 2026). Their individual accuracy is often reported relative to how closely people reproduce their own answers on a retest, and by that measure twins recover most of what is predictable (Park et al., 2026; Toubia et al., 2025), whereas audits find that once item means are removed they explain little of the person-specific variation (Ahn et al., 2026; Wang et al., 2026; Peng et al., 2026). A twin predicts questions that the person was never asked, but in adaptive testing, tutoring, recommendation and longitudinal panels a predictor also sees the person's own earlier behavior and observes each decision after predicting it. This prequential setting (Dawid, 1984), also used in knowledge tracing (Corbett & Anderson, 1995; Piech et al., 2015), is what we call the closed loop, and we use it as the evaluation protocol, Figure 1a. We ask how predictable it makes a person's next choice and which predictors use it. Across two survey panels and two collections of cognitive experiments the answer has one form: most of the signal our predictors extract is the person's most recent relevant behavior, retest agreement measures what one repetition of it reveals rather than what is attainable, and predictors differ in whether they anticipate a departure from it, a changed answer on a survey or a switch in a task.

Twin-2K-500 (Toubia et al., 2025) shows the survey side, Figures 1b, 2a and 2b. Its fourth wave repeats questions that the same 2,058 persons answered in waves 1 to 3. The digital twins released with the dataset do not see these earlier answers, so their fair reference is the item mode of 0.607, and all 13 of them score below it, whereas copying the earlier answer, which a closed-loop predictor sees, reaches 0.712. An LLM given all of the person's earlier answers as a persona matches copying,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

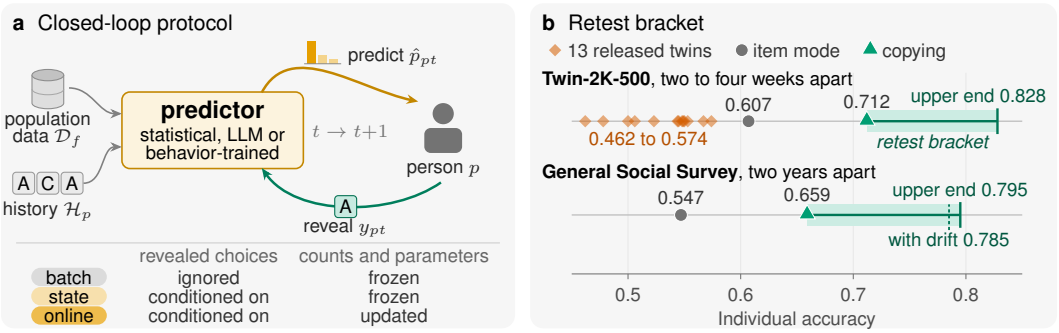


Figure 1: Closed-loop prediction and the retest bracket. Panel a shows the protocol, in which a predictor fitted on the population data $\mathcal{D}_f$ and the history $\mathcal{H}_p$ predicts each test decision of person $p$ before its choice is revealed. Panel b shows individual accuracy, with the bracket of Proposition 1 shaded and, dashed, its upper end with drift on the persons who answered three times, released twins on the rows each covers and the other points on identical rows. The upper ends are estimates under Assumption 1, the dashed one also under one stability shared by the items of a panel, and their intervals cover sampling error only. Panel b covers 2,058 persons and 168,768 retest decisions of Twin-2K-500 and 3,874 persons and 467,793 third answers of the General Social Survey panels.

0.716 against 0.718 on its persons, because it repeats the earlier answer in 98% of rows, and deleting that one line lowers its accuracy to 0.617, 1.3 points above the item mode on the same persons. Its confidence in the earlier answer separates changed from kept answers less well than the item mode does, and within a question its area under the ROC curve is only 0.58, against 0.63 for fitted CAOP-R.

The same data show how to read retest agreement. Retest agreement is the accuracy of copying the earlier answer, so without any assumption it is a lower bound on the accuracy attainable from information that includes that answer. If a person's answers are independent draws from a stable propensity, retest agreement is also the ceiling for a replica that draws each answer from the answer's conditional distribution given what the replica knows, and copying is the nearest-neighbor rule of Cover & Hart (1967), whose inequality bounds the attainable accuracy from above. On Twin-2K-500 this retest bracket runs from 0.712 to 0.828, and dividing by retest agreement credits the item mode with 85%, or 92.6% under the metric of Toubia et al. (2025), more than the 88.4% of their default twin on the same questions. Three waves of the General Social Survey two years apart replicate the pattern, Figure 1b, with copying at 0.659 against 0.547 for the item mode, and a third answer tests the assumption: agreement falls with the time between answers, so on these panels the upper end is an estimate under a model of drift.

Cognitive tasks from Psych-101 and 93 held-out Psych-201 experiments have no retest, but the person's previous choice plays the same role. For statistical predictors, conditioning on it is the largest gain, and they then predict repeats. Reading the person's raw transcript, gemini-3.6-flash is more accurate than the best statistical predictor on both collections and detects more switches at the same false-alarm rate, 78% against 69% for CAOP-sel on Psych-101, Table 1, but a population RNN, a recurrent network trained on other persons' choices and outcomes, matches it there, and on the held-out Psych-201 experiments gemini-3.6-flash is right on switches only slightly more often than the item mode, which ignores the previous choice.

Our contributions are as follows.

- **Protocol and benchmark.** A closed-loop protocol scores statistical, recurrent, LLM and behavior-trained predictors on identical rows of two survey panels and two collections of cognitive experiments and separates the state of the most recent choice from learning.
- **Ceilings.** With the information set explicit, retest agreement is a lower bound on attainable accuracy given the earlier answer and, under stable propensities, a ceiling only for replicas, and the nearest-neighbor inequality gives an upper end that is the accuracy counterpart of the $R^2$ ceiling of Ahn et al. (2026), sharp for binary items given their answer pairs and extended to three occasions with drift.

- **Findings on individuation.** Deleting the line that holds the earlier answer shows that deepseek-flash individuates by lookup, and all three LLMs we ran on Twin-2K-500 repeat the earlier answer when given it. On cognitive tasks the previous choice carries most of the individual signal, a recurrent network trained on other persons matches the best LLM on Psych-101, and on held-out experiments even that LLM is right on switches only slightly more often than the item mode.

## 2 Related work

**LLM digital twins of individuals.** LLMs conditioned on a persona are widely used to simulate survey respondents and experiment participants (Argyle et al., 2023; Park et al., 2026; Toubia et al., 2025; Hu et al., 2026). Twins can reproduce average effects (Wang et al., 2026) but carry little information about which persons deviate from them (Peng et al., 2026; Wang et al., 2026) and miss within-person states (Harry et al., 2026), and once item means are removed their predictions explain only a small share of the person-specific variation on four datasets, Twin-2K-500 among them (Ahn et al., 2026). Persona induction does not improve individual predictions on Psych-201 (Binz et al., 2026), persona twins recover less heterogeneity than a model trained on choices in the same task (Krefeld-Schwalb et al., 2026), a small graph model matches LLM simulators on Twin-2K-500 (Suh et al., 2026), and no LLM reliably beats a persistence baseline in a longitudinal trading study (He et al., 2026), although a respondent's answers to other questions in the prompt do help (Jia et al., 2026; Kinzinger & Hartmann, 2026). These survey evaluations omit copying the earlier answer, and we score LLMs with and without it.

**Behavior-trained models of cognition.** Centaur fine-tunes an LLM on the transcripts of Psych-101 and predicts held-out persons better than domain-specific cognitive models, conditioning each prediction on the person's earlier choices and feedback, and on two experiments it exceeds the estimated noise ceiling because it picks up patterns that depend on context (Binz et al., 2025), so predict-then-observe scoring is not new on these transcripts. Small recurrent networks trained on choices and rewards capture how people learn in such tasks (Dezfouli et al., 2019; Miller et al., 2023; Ji-An et al., 2025), and we train one per experiment on other persons as a baseline. Oh & Gobet (2026) train small language models of the Centaur kind, which we run, fine-tuning on Psych-101 makes Centaur more human-like on held-out Psych-201 tasks (Binz et al., 2026), and whether such models learn the tasks or mainly the sequential structure of choices is debated (Xie & Zhu, 2025; Bowers et al., 2026), although theories derived from choices simulated by Centaur can generalize to held-out human data on multi-attribute decisions (Jagadish et al., 2026). We contribute a common protocol that scores lightweight online predictors, LLMs and behavior-trained models on the same rows.

**Ceilings from repeated measurements.** Digital-twin studies often divide model accuracy by the agreement of respondents with their own answers on a retest (Park et al., 2026; Toubia et al., 2025) or call that agreement an upper bound (Suh et al., 2026). In machine learning, inter-annotator agreement is known not to be the ceiling of classifier performance (Richie et al., 2022), and the Bayes error can be estimated from soft or repeated labels (Ishida et al., 2023; Ushio et al., 2026) or from nearest-neighbor errors (Fukunaga & Hostetler, 1975; Devroye et al., 1996). The classical link is the nearest-neighbor inequality of Cover & Hart (1967), which we apply with the retest answer as the neighbor. Measures of completeness compare the gain of a theory over a naive baseline with that of a flexible predictor trained on the same data (Kleinberg et al., 2017; Fudenberg et al., 2022), as large choice data sets allow (Peterson et al., 2021), whereas the retest bracket bounds attainable accuracy from repeated answers without fitting such a predictor. Noise ceilings in neuroscience bracket the performance of the true model by the agreement of each subject's data with the group average with and without that subject (Nili et al., 2014), whereas our bracket concerns one person's answer and comes from that person's retest. Ahn et al. (2026) use the same pairs on Twin-2K-500 to derive a ceiling on the $R^2$ scale from a variance-components model in the tradition of classical test theory (Lord & Novick, 1968) and generalizability theory (Cronbach et al., 1972), and our bracket is its counterpart for categorical accuracy. Three occasions separate the reliability of an answer from the stability of the attitude (Heise, 1969; Wiley & Wiley, 1970; Alwin, 2007), as Hout & Hastings (2016) show for the General Social Survey, and whether unstable answers come from random responders or from errors around a considered answer is a long-standing question (Zaller & Feldman, 1992).

## 3 CLOSED-LOOP PREDICTION OF INDIVIDUAL DECISIONS

**Task and protocol.** Each person $p$ makes a sequence of decisions $t = 0, \ldots, T_p - 1$. Decision $t$ presents an item $q_{pt}$, the question or situation, with $k_{pt}$ options, and the person's choice is $y_{pt}$. The decisions are split into a history $\mathcal{H}_p$ and a later test segment, and a predictor outputs a probability vector $\hat{p}_{pt} \in \Delta^{k_{pt}-1}$ for every test decision. Persons are assigned to five folds with a fixed seed, and for fold $f$ a predictor is fitted on the population data $\mathcal{D}_f$, every row except the test segments of the persons in fold $f$, Figure 1a. It then receives the history of each person in the fold, and for each test decision in temporal order it outputs $\hat{p}_{pt}$, after which the true choice is revealed. The prediction for decision $t$ thus depends on $\mathcal{D}_f$, $\mathcal{H}_p$ and the test choices before $t$, and on nothing later, and the choice, its label and any feedback text are removed from the row the predictor sees. To separate what the revealed choices contribute, a batch variant of each statistical predictor ignores them, and a state variant of each sequential predictor conditions on them as the most recent choices but keeps every count and parameter fitted on the history, so that the online variant differs from it only by learning. LLMs receive the person's earlier decisions as text and behavior-trained models read the transcript up to the current decision, so all predictor families are scored under the same protocol, except that on Twin-2K-500 the LLM prompts hold waves 1 to 3 only, as a batch variant does.

**Metrics.** The primary metric is individual accuracy, the mean over persons of each person's accuracy on the test segment, which weights every person equally. We also report the log loss of the true choice in bits. Intervals at the 95% level come from a percentile bootstrap over persons with 1000 resamples (Efron & Tibshirani, 1993), and differences between methods from a paired bootstrap on identical rows. These intervals condition on the experiments, so for the main comparisons on cognitive tasks we also report conservative intervals that resample experiments and then persons within them, and sign tests over experiments or studies.

**Datasets.** Table 9 in Appendix D summarizes the datasets, and the appendix gives their construction. *Twin-2K-500* (Toubia et al., 2025) contains four survey waves answered by the same persons one week apart, with wave 4 two weeks after wave 3, and we predict the wave-4 answers to the 108 repeated questions with a single categorical answer from all categorical answers of waves 1 to 3, which include the earlier answer to the same question. Pricing questions show products in a person-specific order at person-specific prices, so we key them by product and price, which yields 508 test items. The *General Social Survey panels* of 2006, 2008 and 2010 (Smith et al., 2017; Hout & Hastings, 2016) interviewed the same persons three times two years apart, and we predict the third answers to 217 attitude, opinion and behavior questions from all answers of the first two waves, Appendix D. *Psych-101* (Binz et al., 2025) consists of transcripts of cognitive experiments, in which each marked choice is a decision whose item is the text presented since the previous feedback. 53 of the 76 experiment files are eligible, and the history is the first 80% of a person's decisions and the test segment the rest. *Psych-201* (Binz et al., 2026) is processed in the same way, keeping only the 188 experiments outside Psych-101, which the behavior-trained models never saw, and 93 of them are eligible. Its subset Psych-201-RT, so named by Oh & Gobet (2026), is their held-out set, of which 13 experiment files are eligible. Statistical predictors are scored on all persons of Table 9, at most 300 per experiment of Psych-101 and 100 of Psych-201, and LLMs on every test item of up to 150 persons of Twin-2K-500 and on the first five test decisions of 10 or 20 persons per experiment of the cognitive tasks, taken in the order of a hash of their identifiers. Behavior-trained models are scored on the first 20 test decisions of held-out persons, and every comparison uses identical rows, Table 8.

## 4 A RETEST BRACKET FOR INDIVIDUAL ACCURACY

Person $p$ answers item $i$ on a first occasion, $Y_{pi}^{(1)}$ in waves 1 to 3, and again on a retest, $Y_{pi}^{(2)}$ in wave 4. A predictor of $Y_{pi}^{(2)}$ sees $X_{pi}$, which contains all earlier answers of the person, including $Y_{pi}^{(1)}$, the retest answers revealed before item $i$ in the closed loop, and data from other persons.

**Assumption 1** (Stable propensity)**.** Each person has a latent state $\theta_p$, and the states and answers of different persons are independent. Given $\theta_p$, all answers of person $p$ are independent, and $Y_{pi}^{(1)}$ and $Y_{pi}^{(2)}$ are both drawn from $\pi_{pi} = \pi_i(\theta_p) \in \Delta^{k_i-1}$.

Let $a_i = \Pr(Y_{pi}^{(1)} = Y_{pi}^{(2)})$ be the retest agreement of item $i$ and $\mathrm{acc}_i^\star = \mathbb{E}\big[\max_j \Pr(Y_{pi}^{(2)} = j \mid X_{pi})\big]$ its Bayes accuracy, the accuracy of the best predictor that uses $X_{pi}$.

**Proposition 1** (Retest bracket)**.** *Under Assumption 1, every item $i$ with $k_i$ options satisfies*

$$a_i \;\le\; \mathrm{acc}_i^\star \;\le\; \mathbb{E}\max_j \pi_{pij} \;\le\; g_{k_i}(a_i) \;\le\; \sqrt{a_i}, \qquad g_k(c) = \tfrac{1}{k}\big(1 + \sqrt{(k-1)(kc-1)}\big).$$

*Proof sketch.* Copying, $\hat{Y} = Y_{pi}^{(1)}$, is feasible and has accuracy $\mathbb{E}\sum_j \pi_{pij}^2 = a_i$. Conditional independence gives $\Pr(Y_{pi}^{(2)} = j \mid X_{pi}) = \mathbb{E}[\pi_{pij} \mid X_{pi}]$, so $\mathrm{acc}_i^\star$ is at most the accuracy $\mathbb{E}\max_j \pi_{pij}$ of an oracle that knows the propensity. For one distribution, $\max_j \pi_j \le g_k(\sum_j \pi_j^2)$, and Jensen's inequality for the concave $g_k$ completes the bound, see Appendix C. □

The earlier answer acts as a nearest neighbor that shares the person's latent state, a sample of size one from the propensity in the sense of Gneiting (2012), so copying is the nearest-neighbor rule and the upper end is the inequality of Cover & Hart (1967), $R \le R^\star\big(2 - kR^\star/(k-1)\big)$ between the error $R = 1 - a_i$ of that rule and the error $R^\star$ of the oracle. The weaker form $\mathrm{acc}_i^\star \le \sqrt{a_i}$ is the categorical counterpart of the classical result that a parallel form correlates with a score as much as its reliability and no predictor more than the square root of it (Spearman, 1904; Lord & Novick, 1968).

**Reading the bracket.** Averaged over each person's test items and then over persons, the bracket on Twin-2K-500 runs from 0.712, the individual accuracy of copying, to 0.828 $[0.824,\ 0.830]$, so dividing by retest agreement places the ceiling at its lower end. Under the assumption, a replica that draws each answer from the conditional distribution $\Pr(Y_{pi}^{(2)} = \cdot \mid X_{pi})$ is right with probability at most $a_i$, attained when it knows the propensity, so retest agreement is the ceiling for such a replica but not for a predictor, whose accuracy the most probable answer maximizes, Remark 1 in Appendix C. The upper end bounds an oracle that knows every propensity, whose accuracy the majority of ever more earlier answers to the item approaches, whereas the best table from item and earlier answer to retest answer reaches 0.719 in sample. For the 465 items with two options the upper end is sharp given the table of answer pairs, because a population in which every person errs around a modal answer at the same rate reproduces each item's table and attains it, whereas on the 393 binary items where copying is at least as accurate as the item mode, persons who either always repeat or answer at random reproduce the table with an oracle no better than copying, two classical readings of unstable answers (Achen, 1975; Converse, 1964). One retest cannot tell them apart, and grouping persons by their agreement with themselves on other items barely lowers the upper end, Proposition 4.

**When the assumption fails.** The lower end needs no assumption. The upper end survives a memory that repeats the earlier answer at a rate that does not depend on the answer and a drift of the propensity that the revealed answers do not predict, with mean zero or toward the answer shares, but it can fail when first answers are noisier than retest answers, when persons repeat some answers more than others, or when the population shifts, Appendix C. The revealed answers of the current wave predict a change, with an odds ratio of 1.14 per 10 points of the share already changed on Twin-2K-500 and 1.15 on the General Social Survey, but using them does not make a rule more accurate, Appendix A. On Twin-2K-500 the answer shares of the two occasions differ at the 5% level for 6% of the 465 binary items, near the nominal rate, but for 47% of the 43 other items, with 35% of rows, so the upper end is least secure where the bracket is widest, and bootstrap intervals miss such failures.

**Three occasions.** A third answer makes the assumption testable, since it implies that the three pairwise agreements are equal, and it separates a drift of propensities toward the answer shares from unreliable answers, as Heise (1969) did for correlations. Under such a drift the agreements in excess of chance identify the stability from the second to the third occasion, and the upper end for the third answer tightens to $g_k\big(a_0 + \kappa_{13}\kappa_{23}/\kappa_{12}\big)$, where $a_0$ is the agreement of two random persons and $\kappa_{uv}$ the agreement of occasions $u$ and $v$ in excess of $a_0$, Proposition 6 in Appendix C.4.

## 5 Predictors

Every predictor returns a smoothed probability vector over the options of the current decision. Appendices E, F and G give exact definitions, hyperparameters, prompts and scoring rules.

**Statistical predictors.** The item mode predicts the item's choice distribution in the population data. Sequential predictors use the person's own choices through option frequencies, the majority of the last five choices, exponentially decayed frequencies, first-order transitions, which we call the Markov predictor, and an ensemble of these four weighted by running log loss, and on the cognitive tasks copying the previous choice predicts it again. A pooled transition predicts from the item and the previous choice with no parameter of the person, and a feedback-conditioned transition also conditions on the feedback that followed that choice, a population version of win-stay lose-shift. Unlike these statistical predictors, a population RNN, one gated recurrent layer trained for each experiment on other persons, reads the item keys, the earlier choices and their coded feedback and learns its own state (Piech et al., 2015; Dezfouli et al., 2019), Appendix A.1. A nearest-neighbor predictor pools the 20 training persons who agree most often with the person. Latent-variable predictors shrink the person's frequencies toward the population, by empirical Bayes shrinkage under a Dirichlet prior or by a person and population mix, or fit a softmax factorization or a nominal-response IRT model (Bock, 1972). On Twin-2K-500, copying puts almost all mass on the earlier answer $Y_{pi}^{(1)}$, and the pooled retest transition estimates $\Pr(Y^{(2)} = j \mid Y^{(1)}, i)$ from other persons' retest answers.

**CAOP.** A lightweight online baseline, cluster-augmented online prediction or CAOP, blends four smoothed distributions over the options, the person's decayed transitions out of the previous choice, the continuations of the two previous choices, the transitions of other persons in the person's cluster, and the item's population distribution, with person weights that grow over the first 15 choices, Appendix E. CAOP-sel keeps these channels but chooses the subset used for each experiment by the log-likelihood of closed-loop predictions on the training persons of each fold, so the choice never sees the held-out persons. For the retest, CAOP-R mixes copying, weighted by how often the person has repeated earlier answers so far, with pooled and within-cluster retest transitions and the item's retest distribution, and fitted CAOP-R cross-fits its weights.

**General LLMs in context.** We prompt deepseek-flash, deepseek-v4-pro, kimi-k2.6 and gemini-3.6-flash with reasoning disabled and take the renormalized probability of the option letters at the first answer token, or a smoothed one-hot answer for gemini-3.6-flash, which returns no log-probabilities. On the cognitive tasks a prompt lists the person's $N$ most recent decisions, revealed test decisions included, or shows the task instructions and the last 4000 characters of the raw transcript, which is what a behavior-trained model reads, and swap controls take the past decisions from another person of the same experiment, with randomized response keys relabeled to the target's roles. On Twin-2K-500 a prompt lists $N$ random earlier answers of the person to other questions, with or without the earlier answer to the target question, or a persona of all categorical answers of waves 1 to 3, with or without the target line, or the released persona text, which omits the repeated questions. We also score the released twins of Toubia et al. (2025), 13 conditions of persona format, prompt and model.

**Behavior-trained models.** Qwentaur models (Oh & Gobet, 2026) are low-rank adapters (Hu et al., 2022) of Qwen3 base models (Yang et al., 2025) trained on Psych-101 like Centaur, and we run two 0.6B adapters with the Qwen3-0.6B base model as control, the one trained on a nested 6.25% subset on Psych-101 persons outside it and the one trained on all of Psych-101 on the held-out Psych-201 experiments of Oh & Gobet (2026). The model reads the person's transcript, so its next-token distribution over the option labels at each of the first 20 test choices is a closed-loop prediction.

## 6 Results

Numbers are individual accuracies unless stated otherwise, with 95% person bootstrap intervals in brackets and paired differences in percentage points, and Appendix B lists the main predictors.

### 6.1 Surveys

**Copying and the retest bracket.** On Twin-2K-500, Figure 1b, the item mode, the most frequent answer to the question, reaches 0.607. Copying the earlier answer reaches 0.712 $[0.708,\ 0.716]$, 10.5 points more, a margin of 10.4 points on the pricing questions, 49% of the rows, whose products, prices and order wave 4 repeats from wave 3, 1.4 on the other binary questions and 14.6 on those

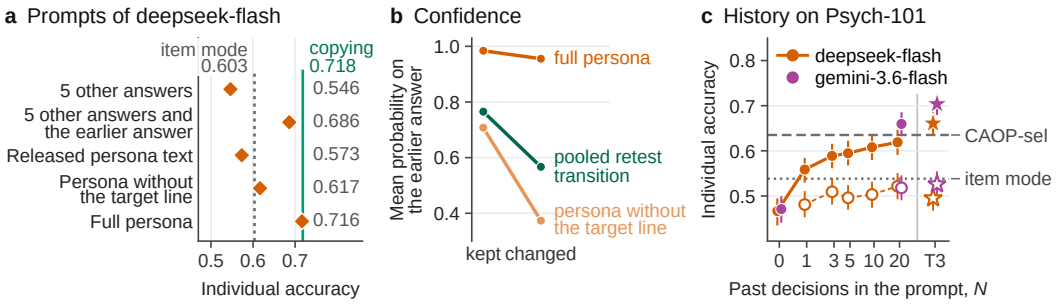


Figure 2: LLMs with and without the person's own behavior. Panels a and b use the 150 persons and 12,316 retest decisions of Twin-2K-500 scored by deepseek-flash, of which the person changed the earlier answer in 3,473, and panel c the 530 persons and 2,611 test decisions of Psych-101 in Figure 3a. In panel c, stars at T3 mark the transcript and open markers another person's history. Error bars are 95% person bootstrap intervals.

with more options, Appendix A.2. The pooled retest transition and fitted CAOP-R add 0.3 and 0.4 points to copying, near the in-sample optimum of 0.719.

**The other answers add little.** Without the earlier answers to the repeated questions, 492 other answers per person lift the softmax factorization 2.7 points $[2.4,\ 3.0]$ above the item mode, 26% of the gain from copying, and stable person-by-item effects dominate the answers, as they do the errors of LLM twins in Ahn et al. (2026), Appendix A.

**Released twins.** All 13 released twins of Toubia et al. (2025), whose personas omit the repeated questions, are significantly below the item mode, the best at 0.574 $[0.571,\ 0.577]$, Table 5 in Appendix B, and under the metric of Toubia et al. (2025) copying is their test-retest accuracy, 0.817, and the twins fall 3.2 to 7.8 points below the item mode, Appendix A.2.

**A second panel over two years.** On the General Social Survey panels copying the answer given two years earlier reaches 0.659 $[0.656,\ 0.661]$ against 0.547 for the item mode, the item mode attains 83% of retest agreement, and the bracket runs to 0.795. The pooled two-answer transition adds 1.2 points to copying, and without the earlier answers the person's other answers recover 50% of the gain from copying, against 26% on Twin-2K-500. A permutation test rejects exchangeable occasions for 44% of the items at the 5% level, and for 70% the agreement four years apart is below both agreements two years apart, against one third under exchangeable occasions, as a drift of the propensities predicts. If the items of each panel share one stability from the second to the third wave, Proposition 6 lowers the upper end on the persons who answered three times from 0.797 to an estimated 0.785 $[0.783,\ 0.787]$, Appendix C.4.

### 6.2 Cognitive Tasks

**The previous choice is the largest gain.** Choices in learning tasks depend on feedback, which violates Assumption 1, so we compare predictors without a bracket. On Psych-101 the item mode reaches 0.532, and copying the previous choice, which is right exactly when the person repeats it, reaches only 0.504. The state variant of CAOP-sel is 6.7 points $[6.1,\ 7.3]$ above the best batch predictor and learning adds 1.4 more, and against each predictor's own batch variant, whose state stays at the last choice of the history, the state variant accounts for 85% to 94% of the gain from the closed loop across the Markov predictor and three variants of CAOP, Figure 5a in Appendix A, a share that this stale state inflates. The pooled transition, which has no parameter of the person, differs from CAOP-sel by $-0.4$ points $[-0.7,\ 0.0]$, whereas the nominal-response IRT model, which has no sequential state, gains only 4.0 points from updating. The individual signal in these tasks thus lies mostly in how the next choice depends on the previous one, as in choice history biases and habits (Urai et al., 2019; Miller et al., 2019), and in line with the low reliability of individual differences in cognitive tasks (Hedge et al., 2018).

Table 1: Individual accuracy on the decisions scored by the LLMs, on repeats of the previous choice and on switches, and the share of switches detected at the rate of predicted switches on repeats of gemini-3.6-flash with the transcript, 17% and 16%, or at their own rate for its variants. The rows are 2,611 decisions of 530 persons in 53 experiments of Psych-101 and 9,140 of 1,860 persons in 93 experiments of Psych-201.

| | Psych-101 | | | | Psych-201, all eligible experiments | | | |
|---|---|---|---|---|---|---|---|---|
| Predictor | All | Repeats | Switches | Detected | All | Repeats | Switches | Detected |
| Item mode | 0.538 | 0.523 | 0.513 | – | 0.579 | 0.582 | 0.574 | – |
| Feedback-conditioned transition | 0.649 | 0.828 | 0.419 | 74% | 0.666 | 0.810 | 0.498 | 68% |
| CAOP-sel | 0.635 | 0.745 | 0.404 | 69% | 0.686 | 0.772 | 0.519 | 71% |
| Population RNN | 0.706 | 0.820 | 0.509 | 79% | 0.684 | 0.755 | 0.552 | 70% |
| deepseek-flash, transcript | 0.661 | 0.762 | 0.487 | 72% | 0.677 | 0.785 | 0.520 | 65% |
| gemini-3.6-flash, transcript | 0.704 | 0.769 | 0.589 | 78% | 0.729 | 0.807 | 0.594 | 74% |
| without the feedback text | 0.677 | 0.758 | 0.547 | 75% | – | – | – | – |
| with shuffled outcomes | 0.653 | 0.741 | 0.505 | 72% | – | – | – | – |

**Behavior-trained models.** On 309 held-out Psych-101 persons, Qwentaur-0.6B trained on the 6.25% subset is 4.7 points [3.4, 6.0] above its base model and 5.1 points [3.0, 7.1] above CAOP-sel, and on shared rows gemini-3.6-flash with the transcript is not significantly more accurate than it once experiments are resampled, Table 3 in Appendix A.

### 6.3 LLMs IN CONTEXT

**Surveys: lookup.** On 150 Twin-2K-500 persons, deepseek-flash given twenty random other answers of the person stays below the item mode, 0.556 against 0.603, and adding the earlier answer to five other answers makes the three models we ran repeat it in 76% to 91% of rows. Given all of the person's categorical answers from waves 1 to 3 as a persona of 18,370 tokens on average, deepseek-flash reaches 0.716 against 0.718 for copying and repeats the earlier answer in 98% of rows, and deleting that single line lowers its accuracy to 0.617, Figure 2a. People keep 72% of these answers, so predicting a repeat is right on most rows, and on the answers they change every predictor that relies on the earlier answer is rarely right, Figure 4 in Appendix A, so the test is whether confidence anticipates a change, Figure 2b. The probability on the earlier answer separates changed from kept answers with an area under the ROC curve of 0.72 [0.71, 0.73] for the LLM, below 0.75 for the item mode and 0.77 [0.76, 0.78] for the pooled retest transition, and within a question the area is 0.58 for the LLM, 0.59 for the transition and 0.63 for fitted CAOP-R, so the separation comes mostly from knowing which questions people change, Appendix A. Neither telling it that about 29% of people change their answer nor reasoning changes this, with the earlier answer repeated in 98% and 99.98% of rows, Appendix F.

**Cognitive tasks: history.** Without history all four LLMs are below the item mode on Psych-101, and the accuracy of deepseek-flash rises from 0.463 to 0.626 with twenty recent decisions and 0.670 with the raw transcript, Figure 2c and Figure 3 in Appendix A. The best LLM condition, gemini-3.6-flash with the transcript, reaches 0.704, 5.4 points [3.8, 7.6] above the best of the 65 statistical predictors, with the winners of both families selected again in every resample, and 6.9 points [4.7, 9.2] above CAOP-sel, [3.0, 11.1] when experiments are also resampled. On all 93 eligible Psych-201 experiments it is 4.3 points [3.1, 5.6] above CAOP-sel, [1.6, 6.7] when experiments are also resampled, and more accurate in 59 experiments and less in 25, whereas deepseek-flash with the transcript differs from CAOP-sel by −0.9 points [−2.2, 0.4].

**Cognitive tasks: switches.** In 54% of the scored decisions of Psych-101 and 53% of those of Psych-201 the person switches away from the previous choice, Table 1. Predictors that condition on the previous choice predict repeats, and the item mode, which ignores it, is right on switches about as often as on repeats. With the transcript, gemini-3.6-flash predicts a switch on 17% of the repeats of Psych-101, its false-alarm rate, as often as CAOP-sel does, and detects 78% of the switches, whereas thresholded to that rate CAOP-sel detects 69%, 9.1 points [5.7, 12.1] fewer, and the feedback-conditioned transition 74%, 4.6 points [1.0, 8.4] fewer, but the population RNN detects 79%, as many. On switches gemini-3.6-flash is 18.6 points [14.7, 22.3] more accurate than CAOP-sel

weighting persons equally, 12.8 weighting decisions, and 7.6 points [3.8, 11.4] more than the item mode, while on repeats it does not differ significantly from CAOP-sel, and the population RNN, which predicts repeats more often, is 5.1 points [1.9, 8.3] more accurate than gemini-3.6-flash on repeats and 8.1 points less on switches, so that overall gemini-3.6-flash differs from it by −0.2 points [−2.1, 1.8]. On all eligible Psych-201 experiments gemini-3.6-flash matches the transition predictors on repeats, −0.3 points [−2.3, 1.6] from the feedback-conditioned transition, and at its rate detects more switches than they and CAOP-sel do, 3.3 points [1.2, 5.8] more than CAOP-sel, but on switches it is only 2.0 points more accurate than the item mode weighting persons equally and 4.8 points weighting decisions, [−2.2, 6.1] and [0.6, 8.7] with experiments resampled, and −0.1 points without the 15 multi-stage experiments, and it is 4.4 points [3.4, 5.6] more accurate than the population RNN overall. Deleting the feedback text from the transcript of Psych-101 lowers it on switches by 4.2 points [1.7, 6.7] and leaves it detecting as many switches as the feedback-conditioned transition, 0.3 points [−3.9, 4.3]. Shuffling the feedback texts among the person's earlier decisions lowers it on switches by 8.5 points [5.9, 11.3], more than deleting them, and its predicted repeats follow the shown outcomes, which the person's choices do not, Appendix A.1.

**How much of the gain is specific to the session.** With the decisions of another person of the same experiment, deepseek-flash loses 71% [60%, 83%] of its gain from twenty decisions on Psych-101 and gemini-3.6-flash 75%, and a swapped transcript leaves both models at or below the item mode, so the gain is specific to the person's session, which includes the outcomes the person experienced.

### 6.4 Where our predictors find individual signal

A decomposition in cross-validated bits measures what each source adds to a pool of the sources before it, so it counts what our predictors extract rather than all the information in a source, Appendix A. On Twin-2K-500 the earlier answer adds 0.238 bits to the item and the answers to other questions then only 0.016, against 0.092 in the other order. On Psych-101 the person, through the nominal-response IRT model, adds 0.116 bits after the item but only 0.028 after the sequence of the person's choices, which itself adds 0.206 after the item, so most of what a person model knows about the next choice is already in the person's recent choices.

## 7 Discussion

**Implications.** Benchmarks of individual prediction that contain repeated measurements should report copying and the retest bracket next to every model, because ratios to retest agreement divide by a lower bound on attainable accuracy and give the item mode 85%. The LLMs we ran individuate by lookup when given the earlier answer, in line with the item-mean surrogacy of Ahn et al. (2026) and the persistence baselines of He et al. (2026), so evaluations of twins should control for the line that holds the answer being predicted. On cognitive tasks the previous choice is the counterpart of that line, and the open problem in both domains is to predict when a person departs from it.

**Limitations.** The upper end rests on Assumption 1 or its drift variant, no predictor we ran exceeds copying by more than 0.4 points on Twin-2K-500 and 1.2 on the General Social Survey, and one retest cannot close the bracket. The LLMs predict the first five test decisions of up to 20 persons per experiment and every test answer of 150 persons on Twin-2K-500, with one prompt per template, one model for the persona ablation and reasoning disabled except in one persona variant and on 106 persons of Psych-101, where it helped on switches, Appendix F. Statistical and recurrent predictors see a stimulus only as an item key, the LLMs may have seen the public transcripts in training, the behavior-trained models are small, folds use one seed, and the intervals on the General Social Survey ignore its sampling design.

**Conclusion.** Individual choices are more predictable from a person's own recent behavior than digital twins suggest, and most of the signal our predictors extract lies in the earlier answer in surveys and the previous choice in a task. Retest agreement is a lower bound on the accuracy attainable given the earlier answer, not a ceiling, and the distance above it lies in the choices where a person departs from that behavior, which remain hard to predict on held-out experiments.

### AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing the text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

All data are public and de-identified: Twin-2K-500 is released under CC-BY-4.0, Psych-101 and Psych-201 under Apache-2.0, and the General Social Survey panel files are distributed publicly for research by NORC. We collected no new data from people, and no LLM received General Social Survey data. The gated Psych-101 test split was not used. Prompts sent to commercial LLM APIs contained only de-identified survey answers and experiment transcripts from these releases. Better individual behavior prediction is dual use. It can support adaptive testing, personalized education and measurement design, and it can also support manipulation or covert profiling. For the second use, our results show that on survey panels, most of the individually predictable signal lies in a person's own earlier answer to the same question, which a predictor can only exploit with repeated access to that person's responses. We did not analyze differences in predictability across demographic subgroups. Such differences can make the same predictor more accurate for some groups than for others, and deployments should measure them.

### REPRODUCIBILITY STATEMENT

The supplementary material contains the complete code. It rebuilds the decision tables from the public releases of Twin-2K-500, the General Social Survey panels, Psych-101 and Psych-201, runs every predictor under the closed-loop protocol with fixed person folds, and writes each reported number to a results file from which the tables and figures are generated. Section 3 and Appendix D describe the datasets and eligibility rules, Appendix E gives all predictor hyperparameters, Appendix F lists every LLM prompt verbatim with API settings, and Appendix C contains the proofs of Proposition 1 and of the propositions on sharpness, strata, drift and three occasions. The row-level predictions of every LLM and behavior-trained condition are included, so every number that involves them can be recomputed without API access or model weights. The behavior-trained models are public LoRA adapters, and the subsets of persons that make one of them a model of held-out persons are reconstructed from the adapter authors' published sampling code (Oh & Gobet, 2026).

## REFERENCES

Christopher H. Achen. Mass Political Attitudes and the Survey Response. *American Political Science Review*, 69(4):1218–1231, 1975. doi: 10.2307/1955282.

Daehwan Ahn, Chengfeng Mao, and Dokyun Lee. Item-Mean Surrogates: Why Richer Persona Data Fail to Improve LLMs as Human Surrogates. *arXiv preprint arXiv:2608.29455*, 2026.

Duane F. Alwin. *Margins of Error: A Study of Reliability in Survey Measurement*. Wiley, Hoboken, NJ, 2007. doi: 10.1002/9780470146316.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao,

Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A Foundation Model to Predict and Capture Human Cognition. *Nature*, 644(8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

Marcel Binz, Elif Akata, Abdullah Almaatouq, Mohammed Alsobay, Oleksii Ariasov, Franziska Brändle, David Broska, Jason W. Burton, Nuno Busch, Frederick Callaway, Vanessa Cheung, Brian Christian, Julian Coda-Forno, Can Demircan, Vittoria Dentella, et al. Post-Training Makes Large Language Models Less Human-Like. *arXiv preprint arXiv:2605.07632*, 2026.

Michael H. Birnbaum. True-and-Error Models Violate Independence and Yet They Are Testable. *Judgment and Decision Making*, 8(6):717–737, 2013. doi: 10.1017/S1930297500004745.

R. Darrell Bock. Estimating Item Parameters and Latent Ability When Responses Are Scored in Two or More Nominal Categories. *Psychometrika*, 37(1):29–51, 1972. doi: 10.1007/BF02291411.

Jeffrey S. Bowers, Guillermo Puebla, Sushrut Thorat, Konstantinos Tsetsos, and Casimir J. H. Ludwig. On the Misuse of Prediction in NeuroAI: A Case Study of Centaur. *PsyArXiv preprint*, 2026. doi: 10.31234/osf.io/v9w37_v5.

Albert H. Bowker. A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, 43(244):572–574, 1948. doi: 10.1080/01621459.1948.10483284.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

Philip E. Converse. The Nature of Belief Systems in Mass Publics. In David E. Apter (ed.), *Ideology and Discontent*, pp. 206–261. Free Press, New York, 1964.

Albert T. Corbett and John R. Anderson. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995. doi: 10.1007/BF01099821.

Thomas M. Cover and Peter E. Hart. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, second edition, 2006.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

A. P. Dawid. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2): 278–292, 1984. doi: 10.2307/2981683.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer, New York, 1996. doi: 10.1007/978-1-4612-0711-5.

Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, and Bernard W. Balleine. Models That Learn How Humans Learn: The Case of Decision-Making and Its Disorders. *PLOS Computational Biology*, 15(6):e1006903, 2019. doi: 10.1371/journal.pcbi.1006903.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the Completeness of Economic Models. *Journal of Political Economy*, 130(4):956–990, 2022. doi: 10.1086/718371.

K. Fukunaga and L. Hostetler. k-Nearest-Neighbor Bayes-Risk Estimation. *IEEE Transactions on Information Theory*, 21(3):285–293, 1975. doi: 10.1109/TIT.1975.1055373.

Tilmann Gneiting. On the Cover-Hart Inequality: What's a Sample of Size One Worth? *Stat*, 1(1): 12–17, 2012. doi: 10.1002/sta4.3.

Tamunotonye Harry, Ivoline C. Ngong, Chima Nweke, Yuanyuan Feng, and Joseph Near. Beyond Fixed Psychological Personas: State Beats Trait, but Language Models are State-Blind. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26440–26468, 2026. doi: 10.18653/v1/2026.findings-acl.1316.

Jiajie He, Yuanhong Jiang, Xintong Chen, Dongling Ni, and Wenjin Liu. Are LLMs Good Financial User Simulators? Multi-view Investor Logic Alignment (MILA). *arXiv preprint arXiv:2609.15727v2*, 2026.

Craig Hedge, Georgina Powell, and Petroc Sumner. The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, 50(3): 1166–1186, 2018. doi: 10.3758/s13428-017-0935-1.

David R. Heise. Separating Reliability and Stability in Test-Retest Correlation. *American Sociological Review*, 34(1):93–101, 1969. doi: 10.2307/2092790.

Michael Hout and Orestes P. Hastings. Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006–2014. *Sociological Science*, 3:971–1002, 2016. doi: 10.15195/v3.a43.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. In *International Conference on Learning Representations*, 2026.

Keito Inoshita. Bias-Corrected Ceilings of Emotion Predictability from Human Label Variation Based on Instance-Level Fano Bounds. *arXiv preprint arXiv:2608.15619*, 2026.

Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the Performance of My Deep Network Too Good to Be True? A Direct Approach to Estimating the Bayes Error in Binary Classification. In *International Conference on Learning Representations*, 2023.

Akshay K. Jagadish, Younes Strittmatter, Nori Jacoby, Eric Schulz, Nathaniel Daw, Thomas L. Griffiths, and Suyog H. Chandramouli. Sparks of In Silico Cognitive Science: Theories from Simulated Data Can Generalize to Humans. *arXiv preprint arXiv:2609.08003*, 2026.

Li Ji-An, Marcus K. Benna, and Marcelo G. Mattar. Discovering Cognitive Strategies with Tiny Recurrent Neural Networks. *Nature*, 644(8078):993–1001, 2025. doi: 10.1038/s41586-025-09142-4.

Mumin Jia, Yilin Chen, Divya Sharma, and Jairo Diaz-Rodriguez. When Can Digital Personas Reliably Approximate Human Survey Findings? *arXiv preprint arXiv:2605.10659*, 2026.

Damjan Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Nikita Karetnikov, Iyad Rahwan, and Davor Svetinovic. Large Language Models as Human Proxies. *Nature Computational Science*, 2026. doi: 10.1038/s43588-026-01060-3.

Leonard Kinzinger and Jochen Hartmann. Synthetic Personalities: How Well Can LLMs Mimic Individual Respondents Using Socio-Economic Microdata? *arXiv preprint arXiv:2606.04592*, 2026.

Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 125–126. ACM, 2017. doi: 10.1145/3033274.3084094.

Ioannis Kontoyiannis, Paul H. Algoet, Yuri M. Suhov, and Abraham J. Wyner. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998. doi: 10.1109/18.669425.

Antonia Krefeld-Schwalb, Eric J. Johnson, Caroline Weaver, and Dirk U. Wulff. Training Data Limits the Prediction of Consumer Heterogeneity by LLM-Based Digital Twins. *PsyArXiv preprint*, 2026. doi: 10.31234/osf.io/97dya_v1.

Vaibhav Kulkarni, Abhijit Mahalunkar, Benoit Garbinato, and John D. Kelleher. Examining the Limits of Predictability of Human Mobility. *Entropy*, 21(4):432, 2019. doi: 10.3390/e21040432.

Dong C. Liu and Jorge Nocedal. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45(1–3):503–528, 1989. doi: 10.1007/BF01589116.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

Xin Lu, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. Approaching the Limit of Predictability in Human Mobility. *Scientific Reports*, 3:2923, 2013. doi: 10.1038/srep02923.

A. E. Maxwell. Comparing the Classification of Subjects by Two Independent Judges. *British Journal of Psychiatry*, 116(535):651–655, 1970. doi: 10.1192/bjp.116.535.651.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Kevin J. Miller, Amitai Shenhav, and Elliot A. Ludvig. Habits Without Values. *Psychological Review*, 126(2):292–311, 2019. doi: 10.1037/rev0000120.

Kevin J. Miller, Maria Eckstein, Matthew M. Botvinick, and Zeb Kurth-Nelson. Cognitive Model Discovery via Disentangled RNNs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 61377–61394, 2023. doi: 10.52202/075280-2682.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

Nick Oh and Fernand Gobet. Small Foundation Models of Human Cognition and Behaviour. In *Conference on Language Modeling*, 2026.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. *arXiv preprint arXiv:2411.10109v3*, 2026. First version 2024, titled Generative Agent Simulations of 1,000 People.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Tianyi Peng, Melanie Brucks, George Gui, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J. Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, Hortense Fong, Elizabeth Friedman, Ariana Guevara, Mohamed Hussein, Kinshuk Jerath, Bruce Kogut, Akshit Kumar, Kristen Lane, Hannah Li, Vicki Morwitz, Oded Netzer, Patryk Perkowski, and Olivier Toubia. Digital Twins Are Funhouse Mirrors: Five Systematic Distortions. *Science Advances*, 12(36): eaeh8260, 2026. doi: 10.1126/sciadv.aeh8260.

Joshua C. Peterson, David D. Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L. Griffiths. Using Large-Scale Experiments and Machine Learning to Discover Theories of Human Decision-Making. *Science*, 372(6547):1209–1214, 2021. doi: 10.1126/science.abe2629.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, volume 28, pp. 505–513, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html.

Alfréd Rényi. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 547–561. University of California Press, 1961.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. Inter-annotator Agreement Is Not the Ceiling of Machine Learning Performance: Evidence from a Comprehensive Set of Simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 275–284, 2022. doi: 10.18653/v1/2022.bionlp-1.26.

Tom W. Smith, Benjamin Schapiro, James Yuli Lee, Eric Milstein, Samantha Neal, Can Ugur, Zhen Xue, and Alice Mingyue Dai. General Social Surveys, 2006–2014 Panel Codebook. NORC at the University of Chicago, National Data Program for the Social Sciences, 2017. URL https://gss.norc.org/get-documentation.html. Codebook for the panels begun in 2006, 2008 and 2010.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010. doi: 10.1126/science.1177170.

Charles Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Alan Stuart. A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification. *Biometrika*, 42(3–4):412–416, 1955. doi: 10.1093/biomet/42.3-4.412.

Joseph Suh, Suhong Moon, and Serina Chang. Graph-Based Alternatives to LLMs for Human Simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46479–46506, 2026. doi: 10.18653/v1/2026.acl-long.2157.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions. *Marketing Science*, 44(6):1446–1455, 2025. doi: 10.1287/mksc.2025.0262.

Anne E. Urai, Jan Willem de Gee, Konstantinos Tsetsos, and Tobias H. Donner. Choice History Biases Subsequent Evidence Accumulation. *eLife*, 8:e46331, 2019. doi: 10.7554/eLife.46331.

Ryota Ushio, Takashi Ishida, and Masashi Sugiyama. Practical Estimation of the Optimal Classification Error with Soft Labels and Calibration. In *International Conference on Learning Representations*, 2026.

Steven Wang, Kyle Hunt, Shaojie Tang, and Kenneth Joseph. When Can LLM Digital Twins Reduce Human Measurement? From Behavioral Fidelity to Statistical Substitutability. *arXiv preprint arXiv:2609.07987*, 2026.

John Robert Warren and Andrew Halpern-Manners. Panel Conditioning in Longitudinal Social Science Surveys. *Sociological Methods & Research*, 41(4):491–534, 2012. doi: 10.1177/0049124112460374.

Matthew D. Webb. Reworking Wild Bootstrap-Based Inference for Clustered Errors. *Canadian Journal of Economics*, 56(3):839–858, 2023. doi: 10.1111/caje.12661.

David E. Wiley and James A. Wiley. The Estimation of Measurement Error in Panel Data. *American Sociological Review*, 35(1):112–117, 1970. doi: 10.2307/2093858.

Hanbo Xie and Jian-Qiao Zhu. Centaur May Have Learned a Shortcut that Explains Away Psychological Tasks. *PsyArXiv preprint*, 2025. doi: 10.31234/osf.io/u7z4t_v2.

En Xu, Yilin Bi, Hongwei Hu, Xin Chen, Zhiwen Yu, Yong Li, Yanqing Hu, and Tao Zhou. Predictability of Complex Systems. *Physics Reports*, 1166:1–107, 2026. doi: 10.1016/j.physrep.2026.01.001.

An Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

John Zaller and Stanley Feldman. A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences. *American Journal of Political Science*, 36(3):579–616, 1992. doi: 10.2307/2111583.

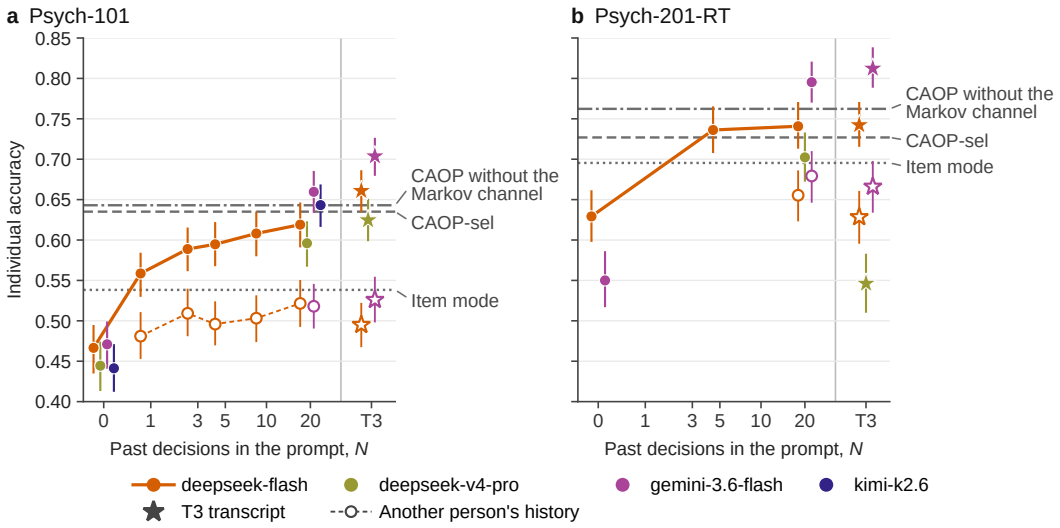


Figure 3: Individual accuracy of LLMs against the number $N$ of the person's recent decisions in the prompt. Panel a covers 530 persons and 2,611 test decisions of Psych-101 scored by every model shown. Panel b covers 260 persons and 1,300 test decisions of Psych-201-RT. Filled circles give the person's own last $N$ decisions in the prompt T1 and filled stars the person's transcript T3. Open markers in the same color give the decisions or the transcript of another person in the same experiment, and a thin dashed line joins them across $N$ where several were run. Grey lines are the item mode, CAOP-sel and CAOP without the Markov channel on the same rows. Error bars are 95% person bootstrap intervals.

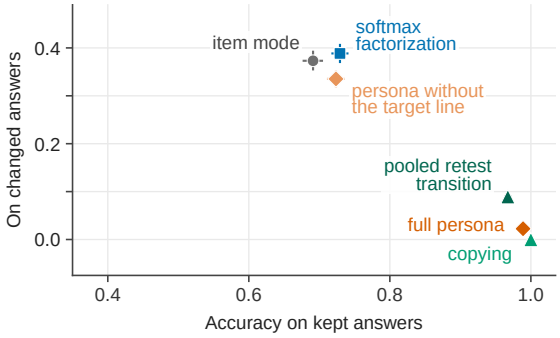


Figure 4: Accuracy on the answers that the person kept and on those the person changed on Twin-2K-500, with the accuracy on kept answers on the horizontal axis. The figure uses the 150 persons and 12,316 retest decisions scored by deepseek-flash, of which the person changed the earlier answer in 3,473. Error bars are 95% person bootstrap intervals.

## A Additional results

Figure 3 shows the accuracy of each LLM against the number of the person's decisions in its prompt on the 530 persons of Psych-101 scored by every model and on Psych-201-RT, and Figure 4 splits the retest decisions of Figure 2a into the answers that the person kept and those the person changed.

**Details of Section 6.** On Psych-101 the best batch predictor is 4.3 points above the item mode. The pooled transition reaches 0.651 against 0.655 for CAOP-sel on Psych-101 and 0.659 against 0.689 on Psych-201, and the nominal-response IRT model gains 3.6 points from updating on Psych-201. With its channels selected on the training persons, CAOP-sel is 4.2 and 3.6 points above that model, the best latent-variable predictor, and has the lowest log loss of any statistical predictor, 0.865 and 0.849 bits, Appendix E. Qwentaur-0.6B reaches 0.706 on its Psych-101 persons, and on 180 persons of

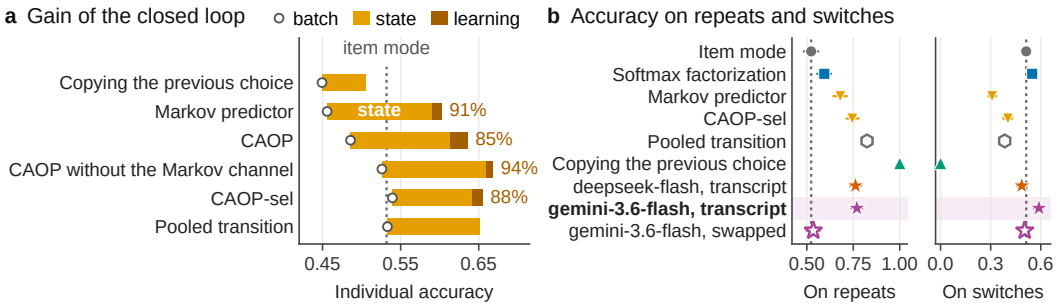


Figure 5: Cognitive tasks on Psych-101. Panel a shows for six predictors the batch variant, the gain of the state variant and the further gain of the online variant, with the share of the state variant in the gain from the predictor's own batch variant to online. Panel b splits the test decisions into repeats of the previous choice and switches, with the item mode dotted. Panel a covers all 7,644 persons of Psych-101 and panel b the 530 persons and 2,611 test decisions scored by every LLM, of which 54% are switches. The row of gemini-3.6-flash with the transcript is shaded. Error bars are 95% person bootstrap intervals.

Table 2: Best predictor of each source in bits per decision. $H_0$ is the log loss of a uniform guess over the options. Sources enter in column order, each column is the reduction in log loss of the best predictor that uses the source relative to the best predictor of the earlier columns, and the residual is the log loss that remains. These increments compare models rather than nest them, and Section 6.4 reports the nested decomposition. On Twin-2K-500 the person column uses the answers to other questions only, and on the cognitive tasks it is the nominal-response IRT model.

| | Bits per decision | | | | |
|---|---|---|---|---|---|
| Dataset | $H_0$ | Item | Person | Earlier answer or sequence | Residual |
| Twin-2K-500 | 1.548 | 0.293 | 0.073 | 0.166 | 1.017 |
| Psych-101 | 1.554 | 0.483 | 0.116 | 0.089 | 0.865 |
| Psych-201 | 1.639 | 0.584 | 0.116 | 0.090 | 0.849 |

Psych-201-RT it reaches 0.745 against 0.709 for its base model and 0.716 for CAOP-sel. Updating changes the accuracy of fitted CAOP-R on Twin-2K-500 by at most 0.07 points. Relative to the upper end of the retest bracket, the best released twin attains 69% and the item mode 73%. On Psych-101 the base model of Qwentaur-0.6B differs from CAOP-sel by 0.4 points $[-1.7,\ 2.4]$. On the rows of Table 3, gemini-3.6-flash is 3.9 points $[1.3,\ 6.6]$ more accurate than Qwentaur-0.6B on Psych-101 and 6.3 points $[3.0,\ 9.6]$ on Psych-201-RT, but these intervals include zero when experiments are also resampled, $[-0.1,\ 8.3]$ and $[-0.4,\ 13.3]$. On Psych-201-RT, CAOP without the Markov channel reaches 0.762 on the rows of gemini-3.6-flash, and the 13 experiment files come from 7 studies. With twenty other answers deepseek-flash is 4.7 points below the item mode on Twin-2K-500, and deleting the target line lowers its accuracy by 10.0 points to 0.617, 1.3 points above the item mode. Without that line, on the 40% of rows where the earlier answer differs from the item mode, the LLM predicts that unseen answer in 32% of rows and the softmax factorization of the other answers in 33%, a difference of $-1.5$ points $[-3.2,\ 0.4]$. The persona without the target line still holds the earlier answers to the other repeated questions, which the softmax factorization does not see, so this comparison favors the LLM. On Psych-101 gemini-3.6-flash with the transcript predicts a switch on 78% of switches and 17% of repeats, against 66% and 17% for CAOP-sel. With the transcript it is more accurate than CAOP-sel in 34 of the 53 experiments of Psych-101 and less accurate in 14. On Psych-201-RT it reaches 0.812 against 0.727 for CAOP-sel, but of the 7 studies 5 favor it, 1 favors CAOP-sel and 1 ties, $p = 0.219$ by a sign test.

**Log loss of copying.** The raw log loss of copying on Twin-2K-500, 2.32 bits, is poor because it puts almost all mass on one option, but after a cross-fitted mix with the uniform distribution it is 1.21 bits against 1.26 for the item mode.

**Where the gain over copying lies on Twin-2K-500.** When persons are grouped by their agreement with themselves on half of their items and scored on the other half, fitted CAOP-R gains 0.7 points $[0.4,\ 1.0]$ over copying in the least consistent quarter and 0.1 points $[-0.1,\ 0.2]$ in the most consistent. The least consistent quarter also holds 41% of the distance from copying to the upper end and the most consistent 14%, but this distance is an item-level bound minus the person's own agreement, so its share grows with inconsistency by construction and does not locate the accuracy of an oracle for any person.

**Confidence and changed answers.** On binary questions the probability that the LLM with the full persona puts on the earlier answer separates changed from kept answers with an area under the ROC curve of 0.60, against 0.71 for the pooled retest transition, and over all questions its area is -0.05 $[-0.06,\ -0.04]$ from that of the transition. Told the base rate of changes, the LLM reaches 0.72 and without the target line 0.74. The run with reasoning returns single answers, so its area of 0.68 reflects only differences between questions. Within a question, the area of the LLM differs from that of the transition by -0.01 $[-0.03,\ 0.00]$. Over all persons of Twin-2K-500 the pooled retest transition reaches 0.77, and on the General Social Survey 0.73, and 0.76 when it conditions on both earlier answers.

**Drift that the current wave predicts.** For every test answer with an earlier answer, a logistic regression cross-fitted over the person folds predicts a change from the change rate of the item and earlier answer among other persons and from the share of changed answers among the person's answers of the current wave revealed before it. The share raises the area under the ROC curve for a change by 0.005 $[0.004,\ 0.006]$ on Twin-2K-500, from 0.77, and by 0.004 $[0.003,\ 0.004]$ on the General Social Survey, from 0.73. In a fit to all rows the share has an odds ratio of 1.14 $[1.13,\ 1.15]$ per 10 points on Twin-2K-500 and 1.15 $[1.14,\ 1.16]$ on the General Social Survey. Predicting the most frequent other answer where it is more likely than a repeat changes the prediction on 6% of the rows of Twin-2K-500 and 8% of those of the General Social Survey, and this rule is 0.3 points $[0.2,\ 0.4]$ and 0.7 points $[0.6,\ 0.8]$ more accurate than copying, against 0.4 and 0.7 without the share, a paired difference of $-0.05$ points $[-0.11,\ 0.02]$ and 0.01 points $[-0.03,\ 0.04]$, so using the drift that the current wave reveals barely changes accuracy.

**Generalizability analysis on Twin-2K-500.** The analysis decomposes the answer variance as in generalizability theory (Cronbach et al., 1972). On all 60 items that every person answered, most of them pricing cells keyed by product, persons account for 5% of the answer variance, items for 12%, the stable person-by-item interaction for 52% and the transient residual for 31%. On the 18 matrix Likert items alone, persons account for 6%, items for 25%, the stable person-by-item interaction for 46% and the transient residual for 22%, so the person-by-item part dominates in both designs. Answers enter both analyses as the position of the chosen option, which suits the ordered Likert and pricing items better than the nominal ones.

**Oracles and sequence-only estimates on the cognitive tasks.** An in-sample per-person transition table on the previous choice reaches 0.700 on Psych-101 and 0.698 on Psych-201. Fano estimates from the entropy rate of a person's choices (Song et al., 2010; Xu et al., 2026) refer to predictors that see only the sequence, and predictors that also see the item are not bounded by them (Inoshita, 2026), since CAOP exceeds the mean Lempel-Ziv estimate (Kontoyiannis et al., 1998) in 21 of 53 experiments on Psych-101 and 44 of 93 on Psych-201, as real predictors exceed such estimates elsewhere (Lu et al., 2013; Kulkarni et al., 2019).

**Swap control and memorization probe.** With the decisions of another person deepseek-flash reaches 0.511 on Psych-101, against 0.626 with twenty of the person's own and 0.463 without history, and a swapped transcript costs 17.9 points for deepseek-flash and 17.8 for gemini-3.6-flash on Psych-101. Continuing 516 verbatim prefixes of the public transcripts, both models reproduced an unpredictable reward or response time in 1.4% to 2.7% of probes, no more often than the most frequent value, 2.7%, but this probe tests no choices, which the shuffled outcomes of Appendix A.1 do. The share of the gain from twenty decisions that is specific to the person is 71% $[60\%,\ 83\%]$ for deepseek-flash and 75% $[61\%,\ 91\%]$ for gemini-3.6-flash on Psych-101, and 47% $[35\%,\ 61\%]$ for gemini-3.6-flash on Psych-201-RT, where another person's history already helps it. With the transcript, a swapped past costs 11.4 points for deepseek-flash and 14.6 for gemini-3.6-flash on

Table 3: Predictors of the cognitive tasks on the decisions scored by both the behavior-trained models and the LLMs. Individual accuracy with 95% person bootstrap intervals. On Psych-101 the behavior-trained model is the adapter trained on the 6.25% subset, evaluated on persons outside it.

| | Psych-101 held-out persons | | Psych-201-RT | |
|---|---|---|---|---|
| Shared with the LLMs | 309 persons, 1,526 decisions | | 180 persons, 900 decisions | |
| Predictor | Accuracy | 95% CI | Accuracy | 95% CI |
| Item mode | 0.531 | [0.491, 0.569] | 0.659 | [0.616, 0.701] |
| Markov predictor | 0.545 | [0.510, 0.579] | 0.552 | [0.516, 0.592] |
| CAOP | 0.581 | [0.545, 0.612] | 0.623 | [0.589, 0.659] |
| CAOP-sel | 0.629 | [0.594, 0.660] | 0.704 | [0.668, 0.742] |
| CAOP without the Markov channel | 0.650 | [0.617, 0.679] | 0.709 | [0.674, 0.749] |
| Qwen3-0.6B base reading the transcript | 0.624 | [0.588, 0.654] | 0.688 | [0.648, 0.724] |
| Qwentaur-0.6B | 0.676 | [0.640, 0.707] | 0.712 | [0.679, 0.747] |
| gemini-3.6-flash with the transcript T3 | 0.714 | [0.681, 0.743] | 0.776 | [0.744, 0.809] |
| deepseek-flash with the transcript T3 | 0.661 | [0.625, 0.692] | 0.717 | [0.679, 0.752] |

Psych-201-RT. Because the swap also replaces the outcomes that the person experienced, specific to the person means specific to the person's own session. The memorization probe draws its 516 prefixes from 11 experiments, its reference rate 2.7% is that of predicting the most frequent value among other persons, and each prefix ends just before a token that it cannot predict, such as a reward or a response time, so the probe tests verbatim recall of the transcripts but not of the choices. On Twin-2K-500, sequence predictors lose 0.231 bits against the item mode because consecutive questions are unrelated.

**Repeats and switches of every predictor.** Of the 94 predictors with predictions on the LLM rows of Psych-101, gemini-3.6-flash with the transcript and the online and batch variants of the softmax factorization and the nearest-neighbor predictor are significantly more accurate than the item mode on switches, the online variants by 3.6 and 3.3 points. The online variants predict a switch on 39% and 42% of repeats, against 17% for gemini-3.6-flash, and are right in only 0.594 and 0.563 of repeats. Qwentaur-0.6B trained on the 6.25% subset differs from the item mode on switches by 1.0 points $[-4.2, 6.5]$ on the rows it shares with the LLMs, and deepseek-flash with the transcript by $-2.6$ points $[-6.6, 1.3]$. The pooled transition is right in 0.824 of repeats and 0.384 of switches. With the transcript deepseek-flash is right in 0.487 of switches. Weighted by decisions instead of persons, the difference between gemini-3.6-flash and CAOP-sel on switches is 12.8 points, since averaging over persons gives a person with one switch the weight of a person with five. On Psych-201-RT, 6 of the experiment files are two-step tasks whose consecutive decisions belong to different stages, and without them gemini-3.6-flash differs from the item mode on switches by 2.2 points $[-4.9, 9.7]$.

**Nested decomposition.** On Twin-2K-500 the answers to other questions add 0.092 bits $[0.087, 0.096]$ to the item and the earlier answer a further 0.155 $[0.150, 0.160]$. Averaged over the two orders, the sequence of the person's choices carries 0.163 bits and the person 0.072 on Psych-101, and 0.163 and 0.071 on Psych-201. The two orders of Section 6.4 end at pools fitted separately, whose log losses differ slightly, so the average over the two orders is not an exact Shapley split. Table 2 gives the best predictor of each source alone.

**Intervals that resample experiments and select again.** Resampling experiments with all of their persons, instead of experiments and then persons, gives slightly narrower intervals, for example $[3.5, 10.3]$ for gemini-3.6-flash with the transcript minus CAOP-sel on Psych-101 and $[4.6, 12.5]$ on Psych-201-RT. Selecting the best of the 18 LLM conditions and the best of the 65 statistical predictors again on every resample gives basic bootstrap intervals of $[3.8, 7.6]$ for the difference of the winners on Psych-101 and $[2.5, 8.9]$ when experiments are resampled with all of their persons, and with 14 and 62 candidates $[2.8, 7.5]$ and $[2.0, 9.1]$ on Psych-201-RT. Over the 7 studies of Psych-201-RT, the sign test in the details of Section 6 is the test we report, and two wild cluster bootstrap tests check it for gemini-3.6-flash minus CAOP-sel, one (Cameron et al., 2008) that enumerates all 128 sign patterns with $p = 0.063$ and one with the six-point weights of Webb (2023) with $p = 0.028$.

**Combining LLMs and statistical predictors.** A log-linear pool of CAOP-sel and deepseek-flash with the transcript, weighted on the training persons of each fold, is 2.9 points [1.8, 4.0] above the more accurate member on Psych-101 and 3.7 points [2.7, 4.6] above CAOP-sel on all eligible Psych-201 experiments. On the Psych-101 rows of deepseek-flash with the transcript, it and CAOP-sel are both right on 52% of the rows, only the LLM on 15%, only CAOP-sel on 12% and neither on 21%, so their errors overlap only partly. This cross-fitted pool gives the LLM an exponent of 0.18. On Twin-2K-500, pooling deepseek-flash with the persona without the target line and the pooled retest transition changes accuracy by −0.7 points [−1.0, −0.4]. Pooled with CAOP-sel on all eligible Psych-201 experiments, gemini-3.6-flash gains 1.5 points [0.9, 2.0], and pooled with CAOP-sel deepseek-flash reaches 0.698 on Psych-101 and 0.722 on Psych-201, with a lower log loss than either member. For gemini-3.6-flash, which returns a single answer, the pool is a mixture with its smoothed one-hot prediction and acts as a gate, and it adds nothing on Psych-101 and Psych-201-RT, −0.8 points [−1.7, 0.1] and −0.1 points [−0.3, 0.2], while lowering the log loss to 1.053 bits on Psych-101. A confidence gate that uses deepseek-flash where its largest probability is high and CAOP elsewhere, Appendix E.5, gains 0.7 points over the LLM alone, [−0.2, 1.5].

**Nested fits.** For CAOP-sel the selection is fully nested: for each outer fold of persons we refit CAOP and its channel variants without the test decisions of that fold, score the persons of the other four folds on their own test decisions and apply to the outer fold the channel subset with the highest likelihood in each experiment. The weights of the log-linear pools in this appendix and the recalibration weights, pooling weights and model choices of the decomposition are also fitted on the persons of the other four folds only, but the statistical predictions of those persons come from models that saw the test decisions of the outer fold. We therefore refitted each pooling weight on the CAOP-sel predictions of the inner runs, which keep only the probability of the observed choice, so the log-linear pools use the binary decisions for this check, and refitted every step of the decomposition on the training folds alone. Against the same binary pools fitted without nesting this changes the pooling weights by at most 0.01, against the reported pools it changes their gains by at most 0.1 points, leaves the model chosen for every source and every increment of the decomposition unchanged to four decimals, and a log-linear pool with its weight fixed in advance at 0.5 still gains 2.3 points [1.6, 3.0] over deepseek-flash on Psych-101 and 2.2 points [1.0, 3.3] over CAOP-sel on Psych-201.

### A.1 Feedback and switches

**The feedback-conditioned transition.** The feedback-conditioned transition extends the pooled transition of Appendix E by the feedback text that followed the previous decision, a population version of win-stay lose-shift without parameters of the person. For each experiment the text is lowercased, every number in it is replaced by its tercile among the first numbers of the feedback texts in the population data, and the rest is kept, so that a text such as *the correct category is P* encodes, with the previous choice, whether that choice was correct. Experiments with more than 50 such codes keep only the tercile of the first number, or a single code if their feedback holds no number, and an experiment with a single code predicts as the pooled transition. For a decision with item $q$ whose person chose $a$ with feedback code $f$ at the decision before, it predicts $\big(n(q,a,f,\cdot)+2P_T\big)/\big(N(q,a,f)+2\big)$, where $P_T$ is the prediction of the pooled transition, and the feedback of the decision being predicted is never used. On the LLM rows of Psych-101 it is right in 0.828 of repeats and 0.419 of switches, 3.5 points [2.0, 5.1] more than the pooled transition on switches, and overall it differs from CAOP-sel by 1.4 points [−0.4, 3.5]. On switches gemini-3.6-flash with the transcript is 17.0 points [13.9, 20.6] more accurate than the feedback-conditioned transition, [11.2, 23.0] when experiments are also resampled, and 18.5 points [14.6, 22.1] in the 42 experiments with trial-by-trial feedback. In the two-step files of Psych-201-RT the sentence that reports the outcome opens the text of the next decision, so there the item key already carries it, and on the LLM rows of Psych-201-RT the feedback-conditioned transition falls back to the pooled transition on 62% of the rows. On Psych-201-RT conditioning on the feedback adds a nonsignificant 1.3 points [−0.2, 3.0] to the pooled transition on switches, and gemini-3.6-flash is 15.1 points [10.8, 19.4] more accurate there than the feedback-conditioned transition. Over all persons of the benchmark, where the feedback cell rather than a fallback makes 80% of the predictions on Psych-101, conditioning on the feedback adds 0.2 points [0.1, 0.3] to the pooled transition on Psych-101 and 1.0 points [0.8, 1.1] on Psych-201, and the feedback-conditioned transition differs from CAOP-sel by −0.2 points [−0.6, 0.2] on Psych-101. Split by feedback, the difference between gemini-3.6-flash and CAOP-sel on switches is 21.3 points in the experiments of Psych-101 with trial-by-trial feedback and 8.0 points [−0.7, 16.6] in the 11 without it, 13.3 points

[3.9, 23.4] less, a contrast that also differs in task. Masking the content of stimuli and feedback removes most of what behavior-trained models learn (Oh & Gobet, 2026), whereas deleting only the feedback text lowers the accuracy of gemini-3.6-flash by 2.7 points and leaves it 13.9 points above the item mode, as the next paragraph shows.

**Transcripts without feedback.** The variant T3nofb shows the same window of the transcript as T3 but deletes the feedback text of every past decision, the text after each choice marker up to the end of its line or the next marker. Outcome statements drop from 54.2 to 7.8 per prompt on Psych-101, but some experiments repeat outcomes elsewhere, such as a running score in the text of the next trial, and 20% of the prompts still hold one. Without the feedback text gemini-3.6-flash reaches 0.677 on Psych-101, 2.7 points [1.2, 4.2] below T3. In two-step tasks whose stages share a line and in one sampling task, the deletion also removes text of the current trial on 3% of the rows, and without these rows the drop is 1.9 points [0.5, 3.4]. On switches gemini-3.6-flash without the feedback text reaches 0.547, 3.4 points [−0.3, 7.0] from the item mode, 12.8 points [9.5, 16.5] above the feedback-conditioned transition and 14.4 points [10.8, 18.0] above CAOP-sel, [8.0, 21.3] when experiments are also resampled, and on repeats it differs from CAOP-sel by 1.3 points [−2.1, 5.0]. Without the feedback text deepseek-flash loses 3.2 points [1.8, 4.6] and reaches 0.457 on switches, against 0.487 with it. On Psych-201-RT, where 38% of the prompts keep an outcome statement, deleting the feedback text does not significantly lower gemini-3.6-flash on switches, 2.2 points [−0.4, 4.9], and leaves it 10.8 points [6.4, 15.4] above CAOP-sel. Most of the advantage over statistical predictors on switches thus survives the deletion, but less than half of that over the item mode does, and outcome statements remain in up to 20% of the prompts, so the outcomes printed in the transcript carry at least part of it.

**Detecting switches.** Over rows, the score one minus the probability of the previous choice separates switches from repeats with an area under the ROC curve of 0.84 [0.82, 0.86] for CAOP-sel, 0.87 [0.85, 0.89] for the feedback-conditioned transition and 0.86 [0.84, 0.88] for deepseek-flash with the transcript on Psych-101. The single answer of gemini-3.6-flash predicts a switch on 17% of repeats and 78% of switches, and at the same rate on repeats the thresholded scores detect 69% of switches for CAOP-sel, whose own answers detect 66%, 74% for the feedback-conditioned transition and 69% for the pooled transition, 9.1 points [5.7, 12.1] and 4.6 points [1.0, 8.4] fewer than gemini-3.6-flash. On Psych-201-RT gemini-3.6-flash predicts a switch on 21% of repeats and 91% of switches, against 85% of switches for CAOP-sel at the same rate. Detecting a switch is necessary but not sufficient for being right on it, since the predictor must also choose among the other options. Without the feedback text gemini-3.6-flash predicts a switch on 18% of repeats and 75% of switches, and at that rate the feedback-conditioned transition detects 75%, a difference of 0.3 points [−3.9, 4.3], so without the feedback text the advantage of gemini-3.6-flash on switches lies in the option it switches to. The 3 multi-stage experiments of Psych-101 contribute switches between stages, and without them gemini-3.6-flash is 7.2 points [3.1, 11.0] above the item mode on switches, [0.8, 13.4] when experiments are also resampled, and 18.9 points [14.8, 23.0] above CAOP-sel.

**Shuffled outcomes.** The variant T3fbshuf keeps the transcript of T3 but moves each feedback text that T3nofb deletes and that ends its line to another past decision of the same person by a seeded derangement, so each moved outcome still appears once. The 516 rows with fewer than two such feedback texts keep the prompt of T3, and outcome statements outside the spans remain in 20% of the prompts. Over the 42 experiments of Psych-101 with trial-by-trial feedback, the shuffle lowers gemini-3.6-flash by 6.3 points [4.6, 8.1], 3.2 points more than deleting the feedback text, and over all LLM rows it lowers it on switches by 8.5 points [5.9, 11.3], [5.0, 12.4] when experiments are also resampled, and 4.2 points [1.8, 7.0] more than the deletion. Pooled over the rows of 26 experiments whose outcomes report points, gemini-3.6-flash with T3 predicts a repeat 26.8 points [20.5, 32.6] more often after a previous outcome above the median than below it, as the persons repeat 25.3 points [18.1, 32.2] more often. With T3fbshuf it repeats 12.9 points [6.1, 19.4] more often after a shown outcome above the median, although the persons' choices do not depend on that shown outcome, 0.3 points [−6.1, 7.1], and 9.9 points [3.6, 16.2] more often after a hidden outcome above the median. The model thus reads the outcomes it is shown, which recalling the person's choices would not produce, although it may also recall part of them, and a shuffled transcript can contradict itself, for example by reporting a correct answer after a different choice. The run cost 2.56 US dollars.

**All eligible Psych-201 experiments.** On the 9,140 rows of the 93 experiments, gemini-3.6-flash with the transcript is 7.5 points [5.7, 9.3] more accurate than CAOP-sel on switches, [3.2, 11.7] when experiments are also resampled, and 9.6 points [7.9, 11.5] more than the feedback-conditioned transition, and on repeats it is 3.5 points [1.5, 5.5] above CAOP-sel and −0.3 points [−2.3, 1.6] from the feedback-conditioned transition. It is more accurate than CAOP-sel on switches in 58 experiments and less in 30, $p = 0.004$ by a sign test, and than the item mode in 48 and less in 41, $p = 0.53$. Weighted by decisions, it is 4.8 points [2.8, 6.8] above the item mode on switches. Weighting persons equally, it is 7.2 points [2.7, 11.9] above the item mode on switches in the 13 experiment files of Psych-201-RT, [−2.1, 15.5] when experiments are also resampled, and 1.1 points [−1.3, 3.6] in the other 80 experiments. At its rate of 16% predicted switches on repeats it detects 74% of switches, against 71% for CAOP-sel, 68% for the feedback-conditioned transition, 66% for the pooled transition and 65% for deepseek-flash with the transcript, and without the 15 multi-stage experiments it detects 7.3 points [4.3, 10.9] more than CAOP-sel. Across all of these rows it is 2.5 points [1.3, 3.6] more accurate than the best statistical predictor, selected again in every resample.

**Population RNN.** For each experiment a gated recurrent network with one layer of 32 units is trained on the persons whose decisions no LLM scores, at most 300 per experiment and a median of 89 on Psych-101 and 69 on Psych-201, on their last 300 decisions, with Adam, early stopping on a tenth of these persons and at most 30 epochs, and it is then run over the whole sequence of each scored person. At each decision it reads embeddings of the item key, the previous choice, the person's previous choice on the same item and the feedback code of the feedback-conditioned transition, together with the standardized first number of that feedback, and a unit test checks that its prediction for a decision does not change when the choice or feedback of that decision or any later decision is altered. Training still improved at the cap of 30 epochs in 34 of the 53 experiments of Psych-101 and 42 of the 93 of Psych-201, so longer training could raise its accuracy, and a first run capped at 8 epochs in most experiments reached only 0.633 on Psych-101. Because it excludes every scored person, it trains on fewer persons than the statistical predictors, on Psych-201 as few as 2 in one experiment. On Psych-101 it reaches 0.706, 0.820 on repeats and 0.509 on switches, with an area under the ROC curve for a switch of 0.90, and thresholded to the rate of gemini-3.6-flash it detects 0.4 points [−3.1, 3.4] more switches than that model. On all eligible Psych-201 experiments it reaches 0.684, 0.755 on repeats and 0.552 on switches, 4.4 points [3.4, 5.6] below gemini-3.6-flash, and at the rate of gemini-3.6-flash it detects 70% of switches, 4.4 points [2.5, 6.8] fewer, with an area of 0.86.

### A.2 TWIN-2K-500 UNDER THE METRIC OF TOUBIA ET AL. (2025)

Toubia et al. (2025) score a prediction of a binary answer by exact match and of an answer on a scale of $k$ ordered options by one minus the absolute difference of the option positions divided by $k - 1$, average within each of 17 tasks and then over tasks, and report 71.72% for their default twin, GPT-4.1-mini with a text persona, 81.72% for the test-retest accuracy of wave 4 against waves 1 to 3, and 87.67% for the ratio of the two averaged over tasks. Their released task-level results reproduce these values to three digits, 0.717, 0.817 and 87.7%. Their questions are the repeated questions of wave 4, and they score 104 of our 108 questions, in 14 of their tasks, leaving out four binary anchoring questions, while all 43 items with more than two options, 35% of the test rows, use ordered scales. Copying the earlier answer is then their test-retest accuracy, 0.817 [0.814, 0.821] on our rows. The item mode reaches 0.756 [0.753, 0.760], 92.6% [92.1%, 93.0%] of test-retest accuracy averaged over tasks, and the item median, the best constant under absolute error, 0.764. Their default twin reaches 0.724 [0.721, 0.727], 88.4% of test-retest accuracy and −3.2 points [−3.6, −2.9] from the item mode, and all 13 released twins fall 3.2 to 7.8 points below the item mode on the rows each covers, each with an interval below zero. They score the twin against the answers of waves 1 to 3 rather than wave 4, which gives 0.722 for the default twin, and random guessing is expected to reach 0.579 against wave 4 on these 14 tasks, against their 59.17% on all 17. For the best released twin under exact match, GPT-4.1 with a JSON persona, their metric gives 0.712 against an exact-match individual accuracy of 0.574, as partial credit for near misses on ordered scales raises its score and averaging over their tasks lowers it, and for their default twin it gives 0.724 against 0.553.

**Question types.** The 440 pricing items, 49% of the test rows, show each person in wave 4 the products, prices and order that the person saw in wave 3, for 100% of the persons and a median of 14

days earlier, so memory can raise copying there. In exact-match individual accuracy, as in the main text, copying exceeds the item mode by 10.4 points on these items, by 1.4 on the 25 other binary items and by 14.6 on the 43 items with more options, which were first answered a median of 27 days before the retest. The upper end is 0.910, 0.858 and 0.701 for the three types, against copying at 0.839, 0.760 and 0.515.

# B Full result tables

Tables 4, 5, 6, 7 and 8 list the predictors behind the figures and most numbers of Section 6.

# C Proofs for Section 4

Throughout, fix an item $i$ with $k = k_i$ options and drop the index $i$, and let $X$ be the information of the predictor, which may include the retest answers to other items revealed before this one. Write $\pi_p = \pi_i(\theta_p)$ for the person's answer distribution, $c_p = \sum_j \pi_{pj}^2$ for its collision probability and $Y^{(1)}, Y^{(2)}$ for the two answers. The Bayes accuracy is the largest accuracy of any predictor $\hat{Y} = f(X)$, because $\Pr(f(X) = Y^{(2)}) = \mathbb{E}\big[\Pr(Y^{(2)} = f(X) \mid X)\big] \le \mathbb{E}\big[\max_j \Pr(Y^{(2)} = j \mid X)\big] = \mathrm{acc}^\star$, with equality for $f(X) \in \arg\max_j \Pr(Y^{(2)} = j \mid X)$.

## C.1 Guessing probability and collision probability

**Lemma 1.** *For $c \in [1/k, 1]$ let*

$$g_k(c) = \frac{1 + \sqrt{(k-1)(kc-1)}}{k}.$$

*Every $\pi \in \Delta^{k-1}$ satisfies $\max_j \pi_j \le g_k\big(\sum_j \pi_j^2\big) \le \big(\sum_j \pi_j^2\big)^{1/2}$. The function $g_k$ is nondecreasing and concave on $[1/k, 1]$, with $g_k(1/k) = 1/k$, $g_k(1) = 1$ and $g_2(c) = \frac{1}{2}\big(1 + \sqrt{2c-1}\big)$. With $e = 1 - \max_j \pi_j$, the first inequality reads $1 - \sum_j \pi_j^2 \le e\,(2 - ke/(k-1))$, the pointwise form of the nearest-neighbor inequality of Cover & Hart (1967).*

*Proof.* Let $m = \max_j \pi_j$ and $c = \sum_j \pi_j^2$, so that $m \ge 1/k$ and $c \ge 1/k$. The other $k-1$ coordinates sum to $1 - m$, so by the Cauchy-Schwarz inequality their squares sum to at least $(1-m)^2/(k-1)$ and

$$c \;\ge\; \varphi(m) := m^2 + \frac{(1-m)^2}{k-1}.$$

Since $\varphi'(m) = 2(km-1)/(k-1) \ge 0$ for $m \ge 1/k$, the function $\varphi$ increases on $[1/k, 1]$ from $1/k$ to 1. The equation $\varphi(m) = c$ is the quadratic $km^2 - 2m + 1 - (k-1)c = 0$, whose root in $[1/k, 1]$ is $m = g_k(c)$, hence $\varphi(m) \le c$ implies $m \le g_k(c)$. Because $\varphi(m) \ge m^2$, every $m$ with $\varphi(m) \le c$ also satisfies $m \le \sqrt{c}$, so $g_k(c) \le \sqrt{c}$. Finally $g_k$ is an affine transformation of the square root of a nonnegative affine function of $c$, hence nondecreasing and concave, and the special values follow by substitution. □

In the terms of Rényi (1961), $H_\infty(\pi) = -\log_2 \max_j \pi_j$ and $H_2(\pi) = -\log_2 \sum_j \pi_j^2$, and the combined inequality $\max_j \pi_j \le (\sum_j \pi_j^2)^{1/2}$ of the lemma is $H_\infty(\pi) \ge \frac{1}{2} H_2(\pi)$.

## C.2 The retest bracket

Figure 6 draws the two ends of the bracket against retest agreement and the bracket of every test item of Twin-2K-500.

**Proposition 2** (Proposition 1 for one item)**.** *Under Assumption 1, $a = \mathbb{E}\, c_p$ and*

$$a \;\le\; \mathrm{acc}^\star \;\le\; \mathbb{E} \max_j \pi_{pj} \;\le\; g_k(a) \;\le\; \sqrt{a}.$$

Table 4: Predictors of the wave-4 answer on Twin-2K-500 under the closed-loop protocol. Individual accuracy with 95% person bootstrap intervals, log loss in bits and individual accuracy by block of questions. Released twins (Toubia et al., 2025) are scored on the rows they cover and LLM prompts on the persons of the LLM sample, with the item mode and copying on the same persons. The negative log-likelihood of LLM rows is after a cross-fitted mix with the uniform distribution, and that of every other row is as predicted. Sequential predictors are omitted, since on a survey their previous choice is the answer to an unrelated question.

| | | All retest items | | | Accuracy by block | | |
|---|---|---|---|---|---|---|---|
| Predictor | Persons | Accuracy | 95% CI | Log loss, bits | Pricing | Likert | Other MC |
| *Population* | | | | | | | |
| Item mode | 2,058 | 0.607 | [0.604, 0.611] | 1.255 | 0.735 | 0.474 | 0.510 |
| Nearest neighbors | 2,058 | 0.629 | [0.625, 0.632] | 1.219 | 0.744 | 0.514 | 0.527 |
| *Latent variable, all wave 1 to 3 answers* | | | | | | | |
| Empirical Bayes shrinkage | 2,058 | 0.714 | [0.710, 0.718] | 1.062 | 0.839 | 0.594 | 0.599 |
| Softmax factorization | 2,058 | 0.658 | [0.655, 0.661] | 1.117 | 0.782 | 0.541 | 0.534 |
| *Answers to other questions only* | | | | | | | |
| Empirical Bayes shrinkage | 2,058 | 0.610 | [0.607, 0.613] | 1.257 | 0.722 | 0.499 | 0.510 |
| Softmax factorization | 2,058 | 0.634 | [0.631, 0.637] | 1.201 | 0.753 | 0.521 | 0.517 |
| *Retest-aware* | | | | | | | |
| Copying | 2,058 | 0.712 | [0.708, 0.716] | 2.320 | 0.839 | 0.586 | 0.602 |
| Pooled retest transition | 2,058 | 0.715 | [0.711, 0.719] | 1.022 | 0.838 | 0.599 | 0.598 |
| CAOP-R fitted | 2,058 | 0.716 | [0.712, 0.719] | 1.017 | 0.838 | 0.597 | 0.602 |
| *Released LLM digital twins* | | | | | | | |
| JSON persona, GPT-4.1 | 2,058 | 0.574 | [0.571, 0.577] | – | 0.717 | 0.442 | 0.425 |
| Text persona, GPT-4.1-mini | 2,058 | 0.553 | [0.550, 0.556] | – | 0.669 | 0.436 | 0.455 |
| Demographics only, GPT-4.1-mini | 2,058 | 0.500 | [0.496, 0.503] | – | 0.588 | 0.422 | 0.396 |
| *deepseek-flash in context, 150 persons of the LLM sample* | | | | | | | |
| Item mode, same rows | 150 | 0.603 | [0.590, 0.616] | 1.264 | 0.732 | 0.473 | 0.498 |
| Copying, same rows | 150 | 0.718 | [0.705, 0.730] | 2.276 | 0.847 | 0.593 | 0.602 |
| T2, $N$=5 | 150 | 0.546 | [0.537, 0.556] | 1.454 | 0.681 | 0.412 | 0.429 |
| T2own, $N$=5 | 150 | 0.686 | [0.674, 0.697] | 1.225 | 0.817 | 0.564 | 0.556 |
| Released persona text | 150 | 0.573 | [0.559, 0.585] | 1.398 | 0.701 | 0.450 | 0.453 |
| Persona without the target line | 150 | 0.617 | [0.606, 0.627] | 1.343 | 0.761 | 0.487 | 0.460 |
| Full persona | 150 | 0.716 | [0.704, 0.728] | 1.185 | 0.847 | 0.588 | 0.601 |
| Full persona with base rate | 150 | 0.716 | [0.703, 0.728] | – | 0.847 | 0.586 | 0.601 |

Table 5: Digital-twin predictions released with Twin-2K-500 (Toubia et al., 2025), scored as predictors of the wave-4 answer on the rows each condition covers. Individual accuracy, the paired difference in individual accuracy to the item mode with its 95% person bootstrap interval, both as fractions, and the share of test rows covered.

| | | | | | Minus item mode | |
|---|---|---|---|---|---|---|
| Persona format | Model | Rows covered, % | Accuracy | 95% CI | Diff. | 95% CI |
| JSON | GPT-4.1 | 100.0 | 0.574 | [0.571, 0.577] | −0.034 | [−0.036, −0.031] |
| JSON, predicted output | GPT-4.1 | 99.1 | 0.567 | [0.564, 0.570] | −0.039 | [−0.042, −0.036] |
| Text | GPT-4.1-mini | 100.0 | 0.553 | [0.550, 0.556] | −0.054 | [−0.058, −0.051] |
| Text | Gemini 2.5 Flash | 100.0 | 0.550 | [0.547, 0.554] | −0.057 | [−0.060, −0.054] |
| Text, repeating questions | GPT-4.1-mini | 100.0 | 0.549 | [0.545, 0.552] | −0.058 | [−0.062, −0.055] |
| Text, default temperature | GPT-4.1-mini | 100.0 | 0.548 | [0.544, 0.551] | −0.060 | [−0.063, −0.056] |
| JSON | GPT-4.1-mini | 48.5 | 0.545 | [0.540, 0.549] | −0.057 | [−0.062, −0.052] |
| Text, reasoning | GPT-4.1-mini | 100.0 | 0.544 | [0.540, 0.547] | −0.064 | [−0.067, −0.060] |
| Summary | GPT-4.1-mini | 99.7 | 0.523 | [0.519, 0.526] | −0.085 | [−0.089, −0.081] |
| JSON, predicted output | GPT-4.1-mini | 90.1 | 0.506 | [0.502, 0.510] | −0.091 | [−0.095, −0.087] |
| Demographics only | GPT-4.1-mini | 100.0 | 0.500 | [0.496, 0.503] | −0.108 | [−0.112, −0.104] |
| Summary and JSON | GPT-4.1-mini | 99.1 | 0.478 | [0.475, 0.482] | −0.131 | [−0.135, −0.126] |
| Fine-tuned on 500 persons | GPT-4.1-mini | 75.7 | 0.462 | [0.457, 0.467] | −0.148 | [−0.154, −0.142] |

Table 6: Predictors of the third answer on the General Social Survey panels of 2006, 2008 and 2010 under the closed-loop protocol. Individual accuracy pooled over the panels with 95% person bootstrap intervals, log loss in bits and individual accuracy by panel. The bracket rows give the best table from item and second answer to third answer, cross-fitted over the folds, the upper end $g_k$ and the upper end with drift on the persons who answered three times, with the stability pooled over the items of each panel and estimated per item, Appendix C.4. The majority of the two earlier answers equals copying under our tie rule and is not listed.

| | Pooled panels | | | Accuracy by panel | | |
|---|---|---|---|---|---|---|
| Predictor | Accuracy | 95% CI | Log loss, bits | 2006 | 2008 | 2010 |
| *Population* | | | | | | |
| Item mode | 0.547 | [0.545, 0.549] | 1.435 | 0.543 | 0.544 | 0.552 |
| *Answers to other questions only* | | | | | | |
| Item mode | 0.547 | [0.544, 0.549] | 1.436 | 0.543 | 0.544 | 0.552 |
| Softmax factorization | 0.602 | [0.600, 0.604] | 1.307 | 0.600 | 0.600 | 0.606 |
| Nominal-response IRT | 0.581 | [0.579, 0.583] | 1.355 | 0.576 | 0.581 | 0.586 |
| *Wave-2 answer to the same question* | | | | | | |
| Copying | 0.659 | [0.656, 0.661] | 2.781 | 0.655 | 0.656 | 0.666 |
| Pooled retest transition | 0.665 | [0.663, 0.668] | 1.198 | 0.661 | 0.663 | 0.671 |
| CAOP-R fitted | 0.666 | [0.663, 0.668] | 1.195 | 0.662 | 0.664 | 0.672 |
| *Wave-1 and wave-2 answers* | | | | | | |
| Pooled two-answer transition | 0.671 | [0.669, 0.673] | 1.163 | 0.665 | 0.671 | 0.677 |
| *Retest bracket* | | | | | | |
| Best table from item and wave-2 answer | 0.665 | [0.663, 0.668] | – | 0.661 | 0.663 | 0.671 |
| Upper end $g_k$ | 0.795 | [0.793, 0.796] | – | 0.792 | 0.793 | 0.800 |
| Upper end with drift, pooled stability | 0.785 | [0.783, 0.787] | – | – | – | – |
| Upper end with drift, per item | 0.783 | [0.781, 0.786] | – | 0.780 | 0.784 | 0.786 |
| *Sample* | | | | | | |
| Persons | 3,874 | | | 1,276 | 1,294 | 1,304 |
| Items | 563 | | | 185 | 179 | 199 |
| Retest pairs | 467,793 | | | 156,890 | 150,576 | 160,327 |

Table 7: Statistical predictors on the Psych-101 benchmark and on the held-out Psych-201 experiments. Individual accuracy with 95% person bootstrap intervals, log loss in bits and the paired difference in individual accuracy to the batch variant that never updates on revealed test decisions.

| | Psych-101, 7,644 persons | | | | Psych-201, 6,974 persons | | | |
|---|---|---|---|---|---|---|---|---|
| Predictor | Accuracy | 95% CI | Log loss, bits | Online minus batch | Accuracy | 95% CI | Log loss, bits | Online minus batch |
| *Population* | | | | | | | | |
| Item mode | 0.532 | [0.526, 0.538] | 1.070 | – | 0.581 | [0.575, 0.587] | 1.056 | – |
| Nearest neighbors | 0.581 | [0.574, 0.587] | 1.035 | 0.006 | 0.615 | [0.609, 0.620] | 1.039 | 0.007 |
| *Sequential online* | | | | | | | | |
| Option frequencies | 0.470 | [0.464, 0.476] | 1.327 | 0.013 | 0.480 | [0.475, 0.485] | 1.512 | 0.007 |
| Majority of last five | 0.506 | [0.500, 0.512] | 1.511 | 0.047 | 0.493 | [0.487, 0.499] | 1.698 | 0.025 |
| Decayed frequencies | 0.517 | [0.511, 0.524] | 1.332 | 0.052 | 0.505 | [0.499, 0.510] | 1.478 | 0.033 |
| Markov predictor | 0.602 | [0.597, 0.608] | 1.047 | 0.146 | 0.597 | [0.592, 0.601] | 1.013 | 0.128 |
| Ensemble of the four | 0.607 | [0.601, 0.612] | 1.037 | – | 0.594 | [0.589, 0.599] | 1.004 | – |
| Copying the previous choice | 0.504 | [0.498, 0.512] | 3.514 | 0.055 | 0.484 | [0.477, 0.490] | 4.620 | 0.031 |
| Pooled transition | 0.651 | [0.646, 0.657] | 0.909 | 0.118 | 0.659 | [0.654, 0.664] | 0.920 | 0.116 |
| Feedback-conditioned transition | 0.653 | [0.647, 0.659] | 0.885 | – | 0.669 | [0.664, 0.674] | 0.901 | – |
| CAOP | 0.635 | [0.631, 0.641] | 0.985 | 0.150 | 0.650 | [0.645, 0.654] | 0.954 | 0.145 |
| *Latent variable* | | | | | | | | |
| Empirical Bayes shrinkage | 0.585 | [0.579, 0.591] | 0.997 | 0.012 | 0.616 | [0.611, 0.622] | 1.005 | 0.009 |
| Person and population mix | 0.582 | [0.576, 0.588] | 1.014 | 0.011 | 0.623 | [0.618, 0.629] | 1.008 | 0.012 |
| Softmax factorization | 0.612 | [0.607, 0.618] | 0.956 | 0.040 | 0.650 | [0.644, 0.655] | 0.945 | 0.033 |
| Nominal-response IRT | 0.613 | [0.608, 0.619] | 0.954 | 0.040 | 0.653 | [0.648, 0.658] | 0.939 | 0.036 |
| *CAOP with selected channels* | | | | | | | | |
| CAOP without the Markov channel | 0.668 | [0.663, 0.673] | 0.979 | 0.142 | 0.693 | [0.688, 0.697] | 0.953 | 0.139 |
| CAOP-sel | 0.655 | [0.650, 0.660] | 0.865 | 0.116 | 0.689 | [0.684, 0.694] | 0.849 | 0.114 |
| CAOP-sel by accuracy | 0.671 | [0.666, 0.675] | 0.918 | – | 0.697 | [0.692, 0.702] | 0.910 | – |

Table 8: LLM and behavior-trained predictors, each on the rows of its sample, with paired differences in individual accuracy, in percentage points, to the item mode and to CAOP-sel, or to fitted CAOP-R on Twin-2K-500, on identical rows. Templates are defined in Section 5 and Appendix F, and $N$ is the number of the person's earlier decisions in the prompt.

| | | | Difference in percentage points | | | |
|---|---|---|---|---|---|---|
| | | | Minus item mode | | Minus CAOP-sel | |
| Model | Prompt | Accuracy | Diff. | 95% CI | Diff. | 95% CI |
| *Psych-101, 1,060 persons of the LLM sample* | | | | | | |
| deepseek-flash | T1, $N$=0 | 0.463 | −7.2 | [−9.7, −4.8] | −18.3 | [−20.5, −16.0] |
| deepseek-flash | T1, $N$=1 | 0.568 | 3.2 | [0.4, 5.8] | −7.9 | [−9.7, −6.1] |
| deepseek-flash | T1, $N$=3 | 0.597 | 6.2 | [3.8, 8.6] | −5.0 | [−6.6, −3.4] |
| deepseek-flash | T1, $N$=5 | 0.606 | 7.0 | [4.6, 9.5] | −4.1 | [−5.6, −2.3] |
| deepseek-flash | T1, $N$=10 | 0.612 | 7.7 | [5.2, 10.3] | −3.4 | [−5.3, −1.6] |
| deepseek-flash | T1, $N$=20 | 0.626 | 9.1 | [6.6, 11.4] | −2.0 | [−3.8, −0.3] |
| deepseek-flash | T1fb, $N$=20 | 0.627 | 9.1 | [6.6, 11.5] | −2.0 | [−3.7, −0.2] |
| deepseek-flash | T1swap, $N$=1 | 0.473 | −6.2 | [−8.7, −4.0] | −17.3 | [−19.5, −15.3] |
| deepseek-flash | T1swap, $N$=3 | 0.487 | −4.8 | [−7.3, −2.6] | −15.9 | [−18.1, −13.9] |
| deepseek-flash | T1swap, $N$=5 | 0.488 | −4.8 | [−7.2, −2.5] | −15.9 | [−18.0, −13.9] |
| deepseek-flash | T1swap, $N$=10 | 0.492 | −4.3 | [−6.9, −2.0] | −15.4 | [−17.6, −13.4] |
| deepseek-flash | T1swap, $N$=20 | 0.511 | −2.5 | [−4.8, −0.2] | −13.6 | [−15.7, −11.6] |
| deepseek-flash | T3 | 0.670 | 13.4 | [11.0, 15.7] | 2.3 | [0.6, 4.0] |
| deepseek-flash | T3swap | 0.490 | −4.5 | [−7.0, −2.2] | −15.6 | [−17.9, −13.5] |
| deepseek-flash | T3nofb | 0.635 | 10.0 | [7.5, 12.3] | −1.1 | [−2.9, 0.5] |
| *Psych-101, 530 persons of the LLM sample* | | | | | | |
| deepseek-v4-pro | T1, $N$=0 | 0.444 | −9.4 | [−13.1, −5.9] | −19.1 | [−22.7, −15.5] |
| deepseek-v4-pro | T1, $N$=20 | 0.596 | 5.8 | [2.3, 9.3] | −3.9 | [−6.4, −1.2] |
| deepseek-v4-pro | T3 | 0.625 | 8.6 | [5.6, 11.8] | −1.0 | [−3.4, 1.3] |
| gemini-3.6-flash | T1, $N$=0 | 0.471 | −6.7 | [−10.0, −3.5] | −16.4 | [−19.8, −12.7] |
| gemini-3.6-flash | T1, $N$=20 | 0.660 | 12.1 | [9.2, 15.2] | 2.5 | [0.1, 4.9] |
| gemini-3.6-flash | T1swap, $N$=20 | 0.518 | −2.0 | [−5.2, 1.2] | −11.7 | [−14.8, −8.6] |
| gemini-3.6-flash | T3 | 0.704 | 16.5 | [13.6, 19.7] | 6.9 | [4.7, 9.2] |
| gemini-3.6-flash | T3swap | 0.526 | −1.2 | [−4.6, 2.1] | −10.9 | [−13.9, −7.7] |
| gemini-3.6-flash | T3nofb | 0.677 | 13.9 | [10.9, 17.0] | 4.2 | [2.1, 6.5] |
| kimi-k2.6 | T1, $N$=0 | 0.441 | −9.7 | [−13.1, −6.3] | −19.4 | [−22.8, −15.7] |
| kimi-k2.6 | T1, $N$=20 | 0.643 | 10.5 | [7.6, 13.7] | 0.8 | [−1.4, 3.1] |
| *Psych-201-RT, 260 persons* | | | | | | |
| deepseek-flash | T1, $N$=0 | 0.629 | −6.6 | [−9.9, −3.2] | −9.8 | [−13.5, −6.1] |
| deepseek-flash | T1, $N$=5 | 0.736 | 4.1 | [0.5, 7.6] | 0.9 | [−2.2, 4.1] |
| deepseek-flash | T1, $N$=20 | 0.741 | 4.5 | [1.2, 8.2] | 1.4 | [−1.8, 4.8] |
| deepseek-flash | T1swap, $N$=20 | 0.655 | −4.0 | [−7.7, −0.4] | −7.2 | [−10.8, −3.7] |
| deepseek-flash | T3 | 0.742 | 4.7 | [0.8, 8.9] | 1.5 | [−1.7, 5.1] |
| deepseek-flash | T3c | 0.758 | 6.2 | [2.5, 10.2] | 3.1 | [0.1, 6.2] |
| deepseek-flash | T3swap | 0.628 | −6.7 | [−10.8, −2.7] | −9.8 | [−13.7, −6.1] |
| deepseek-flash | T3nofb | 0.722 | 2.7 | [−1.4, 7.0] | −0.5 | [−3.8, 3.0] |
| deepseek-v4-pro | T1, $N$=20 | 0.702 | 0.7 | [−2.7, 4.3] | −2.5 | [−5.5, 0.8] |
| deepseek-v4-pro | T3 | 0.546 | −14.9 | [−20.3, −9.8] | −18.1 | [−23.0, −13.2] |
| deepseek-v4-pro | T3c | 0.594 | −10.2 | [−13.9, −6.3] | −13.3 | [−16.7, −9.8] |
| gemini-3.6-flash | T1, $N$=0 | 0.550 | −14.5 | [−18.2, −10.5] | −17.7 | [−21.9, −13.7] |
| gemini-3.6-flash | T1, $N$=20 | 0.795 | 10.0 | [6.7, 13.8] | 6.8 | [4.0, 10.0] |
| gemini-3.6-flash | T1swap, $N$=20 | 0.679 | −1.6 | [−5.4, 2.2] | −4.8 | [−8.5, −1.2] |
| gemini-3.6-flash | T3 | 0.812 | 11.7 | [8.2, 15.5] | 8.5 | [5.3, 11.9] |
| gemini-3.6-flash | T3c | 0.808 | 11.3 | [8.0, 15.2] | 8.2 | [5.1, 11.2] |
| gemini-3.6-flash | T3swap | 0.666 | −2.9 | [−6.8, 1.0] | −6.1 | [−10.2, −2.2] |
| gemini-3.6-flash | T3nofb | 0.797 | 10.2 | [6.8, 13.7] | 7.0 | [3.6, 10.2] |
| *Psych-201, 1,860 persons of the LLM sample of all experiments* | | | | | | |
| deepseek-flash | T1, $N$=20 | 0.667 | 8.8 | [7.2, 10.4] | −1.8 | [−3.1, −0.6] |
| deepseek-flash | T3 | 0.677 | 9.8 | [8.2, 11.5] | −0.9 | [−2.2, 0.4] |
| gemini-3.6-flash | T3 | 0.729 | 15.0 | [13.4, 16.7] | 4.3 | [3.1, 5.6] |
| *Twin-2K-500, 150 persons of the LLM sample, CAOP-R fitted in place of CAOP-sel* | | | | | | |
| deepseek-flash | T2, $N$=0 | 0.481 | −12.3 | [−13.7, −10.7] | −24.1 | [−25.8, −22.4] |
| deepseek-flash | T2, $N$=5 | 0.546 | −5.7 | [−6.7, −4.6] | −17.5 | [−18.7, −16.2] |
| deepseek-flash | T2, $N$=20 | 0.556 | −4.7 | [−5.8, −3.7] | −16.5 | [−17.7, −15.2] |
| deepseek-flash | T2own, $N$=5 | 0.686 | 8.3 | [7.0, 9.4] | −3.5 | [−4.0, −2.9] |
| deepseek-flash | released persona text | 0.573 | −3.0 | [−4.0, −1.9] | −14.8 | [−15.9, −13.6] |
| deepseek-flash | persona without the target line | 0.617 | 1.3 | [0.2, 2.5] | −10.5 | [−11.6, −9.3] |
| deepseek-flash | full persona | 0.716 | 11.3 | [9.9, 12.6] | −0.5 | [−0.9, −0.1] |
| deepseek-flash | full persona with base rate | 0.716 | 11.2 | [9.8, 12.6] | −0.5 | [−0.9, −0.2] |
| *Twin-2K-500, 100 persons of the LLM sample, CAOP-R fitted in place of CAOP-sel* | | | | | | |
| deepseek-v4-pro | T2, $N$=5 | 0.481 | −11.8 | [−13.3, −10.2] | −23.9 | [−25.5, −22.2] |
| deepseek-v4-pro | T2own, $N$=5 | 0.628 | 2.9 | [1.3, 4.3] | −9.2 | [−10.2, −8.3] |
| gemini-3.6-flash | T2, $N$=5 | 0.559 | −4.0 | [−5.2, −2.8] | −16.1 | [−17.9, −14.2] |
| gemini-3.6-flash | T2own, $N$=5 | 0.690 | 9.1 | [7.6, 10.6] | −3.0 | [−3.7, −2.3] |
| *Twin-2K-500, 50 persons of the LLM sample, CAOP-R fitted in place of CAOP-sel* | | | | | | |
| deepseek-flash with reasoning | full persona | 0.710 | 10.7 | [8.3, 13.3] | −0.1 | [−0.8, 0.5] |
| *Behavior-trained LMs, Psych-101, 309 held-out persons* | | | | | | |
| Qwen3-0.6B base | transcript | 0.659 | 13.0 | [9.6, 16.4] | 0.4 | [−1.7, 2.4] |
| Qwentaur-0.6B-f0.0625 | transcript | 0.706 | 17.7 | [14.3, 20.8] | 5.1 | [3.0, 7.1] |
| *Behavior-trained LMs, Psych-201-RT, 180 persons* | | | | | | |
| Qwen3-0.6B base | transcript | 0.709 | 4.4 | [1.5, 7.8] | −0.7 | [−2.9, 1.5] |
| Qwentaur-0.6B | transcript | 0.745 | 8.0 | [5.1, 11.3] | 2.9 | [0.7, 5.1] |

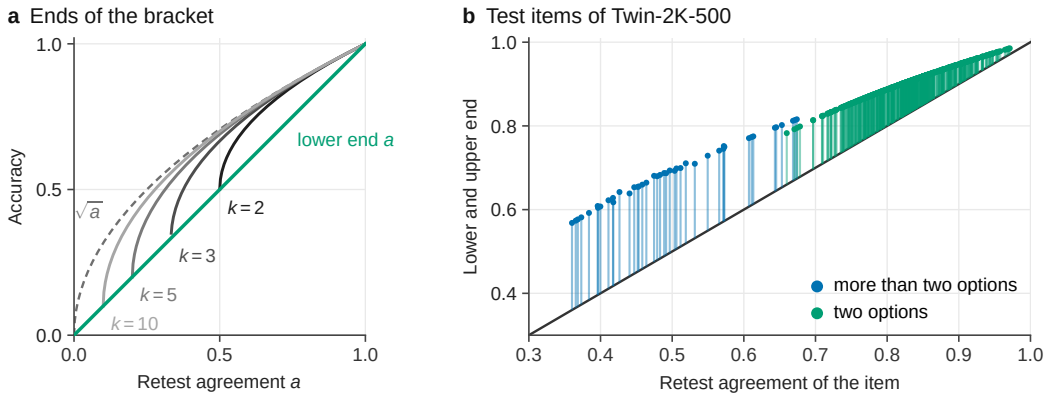


Figure 6: The retest bracket of Proposition 1. Panel a shows the lower end $a$, the upper end $g_k(a)$ for $k = 2$, 3, 5 and 10 options, which starts on the lower end at $a = 1/k$, and the weaker bound $\sqrt{a}$. Panel b shows each test item of Twin-2K-500 as a segment from its retest agreement to its upper end. Of the 508 test items, 465 have two options.

*Proof.* Given $\theta_p$, the answers $Y^{(1)}$ and $Y^{(2)}$ are independent with the same law $\pi_p$, so $\Pr(Y^{(1)} = Y^{(2)} \mid \theta_p) = \sum_j \pi_{pj}^2 = c_p$ and $a = \mathbb{E}\, c_p$. In particular $a \geq 1/k$, so $g_k(a)$ is defined.

Lower end. The earlier answer $Y^{(1)}$ is part of $X$, so copying, $\hat{Y} = Y^{(1)}$, is a predictor of the form $f(X)$. Its accuracy is $\Pr(Y^{(1)} = Y^{(2)}) = a$, and the Bayes accuracy is at least the accuracy of any such predictor.

Upper end. Under Assumption 1, $X$ consists of other answers of the same person, which are independent of $Y^{(2)}$ given $\theta_p$, and of data from other persons, which are independent of person $p$. Conditional independence of $Y^{(2)}$ and $X$ given $\theta_p$ therefore gives

$$\Pr(Y^{(2)} = j \mid X) = \mathbb{E}\big[\Pr(Y^{(2)} = j \mid \theta_p, X) \,\big|\, X\big] = \mathbb{E}\big[\pi_{pj} \,\big|\, X\big].$$

A maximum of conditional expectations is at most the conditional expectation of the maximum, so $\max_j \Pr(Y^{(2)} = j \mid X) \leq \mathbb{E}[\max_j \pi_{pj} \mid X]$ and, taking expectations, $\mathrm{acc}^\star \leq \mathbb{E} \max_j \pi_{pj}$. Lemma 1 and Jensen's inequality for the concave $g_k$ give $\mathbb{E} \max_j \pi_{pj} \leq \mathbb{E}\, g_k(c_p) \leq g_k(\mathbb{E}\, c_p) = g_k(a)$, and $g_k(a) \leq \sqrt{a}$ by the lemma. □

*Remark* 1 (Replicas and repeated answers). Under Assumption 1, a replica that draws $\hat{Y}$ from $\Pr(Y^{(2)} = \cdot \mid X)$, independently of $Y^{(2)}$ given $X$, is right with probability $\mathbb{E} \sum_j \Pr(Y^{(2)} = j \mid X)^2 \leq a$, with equality when $X$ determines $\pi_p$, whereas the Bayes accuracy is at least $a$ when $X$ contains $Y^{(1)}$. If the person answered the item $m$ times before the retest, each answer drawn from $\pi_p$ independently given $\theta_p$, the majority of these answers, under any rule for ties, is right with probability $a$ for $m = 1$, where it is copying, and with a probability that tends to $\mathbb{E} \max_j \pi_{pj}$ as $m$ grows.

*Proof.* The replica is right with probability $\mathbb{E} \sum_j \Pr(\hat{Y} = j \mid X) \Pr(Y^{(2)} = j \mid X) = \mathbb{E} \sum_j \Pr(Y^{(2)} = j \mid X)^2$. By the conditional independence in the proof of Proposition 2, $\Pr(Y^{(2)} = j \mid X) = \mathbb{E}[\pi_{pj} \mid X]$, and Jensen's inequality gives $\mathbb{E}[\pi_{pj} \mid X]^2 \leq \mathbb{E}[\pi_{pj}^2 \mid X]$, whose sum over $j$ has expectation $\mathbb{E}\, c_p = a$, with equality when $\pi_p$ is a function of $X$. Given $\theta_p$ the $m$ earlier answers are independent draws from $\pi_p$, so their shares converge to $\pi_p$ almost surely, every option with the largest share eventually has the largest propensity, the retest answer is independent of the earlier answers given $\theta_p$, and bounded convergence gives the limit. □

Under Assumption 1, retest agreement thus bounds the accuracy of a replica from above and that of prediction from the earlier answer from below, and the accuracy of the oracle is the limit of asking the same item ever more often.

**From items to individual accuracy.** Individual accuracy averages, over persons, the accuracy over each person's test items $I_p$. Suppose that which items a person answers depends on the person only through $\theta_p$, as when alternative versions of a question are assigned at random or when a person's own refusals and absences decide which answers exist. With the weights $w_{pi} = \mathbf{1}[i \in I_p]/|I_p|$ of a row in individual accuracy and the weighted agreement $a_i^w = \mathbb{E}[w_{pi}c_{pi}]/\mathbb{E}[w_{pi}]$, the argument of Proposition 2 and Jensen's inequality under these weights bound the expected individual accuracy of any predictor by $\mathbb{E}_p\big[|I_p|^{-1}\sum_{i\in I_p} g_{k_i}(a_i^w)\big]$, which copying attains with $a_i^w$ in place of $g_{k_i}(a_i^w)$, and $a_i^w = a_i$ when $I_p$ is independent of $\theta_p$. On the General Social Survey, where refusals, answers of don't know and attrition make $I_p$ depend on the person, the weighted agreement changes the upper end by 0.03 points, to 0.794. We estimate $a_i$ by the empirical agreement on item $i$ and report bootstrap intervals over persons for the resulting bracket. Because $g_k$ is concave, the plug-in value $g_{k_i}(\hat{a}_i)$ is biased slightly downward, by an amount that shrinks with the number of pairs per item. On Twin-2K-500 this upper end is 0.828 $[0.824,\ 0.830]$ on the scale of individual accuracy, 11.6 percentage points $[11.4,\ 11.7]$ above copying, and the looser $\sqrt{a_i}$ gives 0.837.

**Sharpness and strata.** The upper end cannot be improved from the table of answer pairs alone when the item has two options.

**Proposition 3** (Sharpness for two options)**.** *For an item with two options and answer shares $\mu$, every table of answer pairs that Assumption 1 can produce is produced by a distribution of propensities with $\mathbb{E}\max_j \pi_{pj} = g_2(a)$ and by one with $\mathbb{E}\max_j \pi_{pj} = \max\{a, \max_j \mu_j\}$, and every distribution of propensities that produces it has $\max\{a, \max_j \mu_j\} \le \mathbb{E}\max_j \pi_{pj} \le g_2(a)$.*

*Proof.* Write $\pi_p = (q_p, 1-q_p)$ and $D_p = q_p - \frac{1}{2}$, so that $\max_j \pi_{pj} = \frac{1}{2} + |D_p|$ and $c_p = \frac{1}{2} + 2D_p^2$. The table of pairs is symmetric and determined by $a = \frac{1}{2} + 2\mathbb{E}D_p^2$ and the share $\mu_1 = \frac{1}{2} + \mathbb{E}D_p$ of the first option, and we label the options so that $\mu_1 = \max_j \mu_j$. Let $d = \sqrt{2a-1}/2$, so that $\mathbb{E}D_p^2 = d^2$, $|\mu_1 - \frac{1}{2}| \le d$ by Jensen's inequality, and $\mathbb{E}\max_j \pi_{pj} \le \frac{1}{2} + d = g_2(a)$. If $d = 0$ every propensity is uniform and there is nothing to show. Otherwise the distribution with $D_p = d$ with probability $\frac{1}{2}\big(1 + (\mu_1 - \frac{1}{2})/d\big)$ and $D_p = -d$ with the remaining probability reproduces $a$ and $\mu_1$ and attains $\frac{1}{2} + d$. For the lower end, $|D_p| \le \frac{1}{2}$ gives $2D_p^2 \le |D_p|$, so $\mathbb{E}|D_p| \ge \max\{|\mathbb{E}D_p|, 2\mathbb{E}D_p^2\}$, which is $\mathbb{E}\max_j \pi_{pj} \ge \max\{\mu_1, a\}$. Let $m = \mu_1 - \frac{1}{2} \ge 0$. If $a \ge \mu_1$, the distribution with $D_p = \frac{1}{2}, 0$ and $-\frac{1}{2}$ with probabilities $a + \mu_1 - 1$, $2 - 2a$ and $a - \mu_1$ reproduces $a$ and $\mu_1$ and attains $a$. If $a < \mu_1$, then $m^2 \le d^2 < m/2$, and the distribution with $D_p = m$ with probability $1 - \lambda$ and otherwise $D_p \in \{0, \frac{1}{2}\}$ with mean $m$, where $\lambda = (d^2 - m^2)/(m/2 - m^2)$, reproduces $a$ and $\mu_1$ and attains $\mu_1$. □

The distribution that attains the upper end gives every person the same rate of error around a modal answer, close to the measurement-error reading of unstable survey answers of Achen (1975), and when every person either always gives the same answer or answers at random the oracle is no more accurate than copying, the reading of Converse (1964), which reproduces the table only when copying is at least as accurate as the item mode, as on 393 of the 465 binary items of Twin-2K-500. Where it does, which of the two holds cannot be decided from one retest. Both ends can be written as true-and-error models of repeated choices (Birnbaum, 2013), the upper end as the simplest one, with an error rate common to all persons. Information about the person from other items can, however, sharpen the upper end.

**Proposition 4** (Stratified upper end)**.** *Let $S$ be a function of the person's answers to items other than $i$, on either occasion, and $a_i(s) = \Pr(Y_{pi}^{(1)} = Y_{pi}^{(2)} \mid S = s)$. Under Assumption 1, $\mathrm{acc}_i^\star \le \mathbb{E}\, g_{k_i}\big(a_i(S)\big) \le g_{k_i}(a_i)$.*

*Proof.* Given $\theta_p$ the two answers to item $i$ are independent of $S$, so $a_i(s) = \mathbb{E}[c_p \mid S = s]$. Lemma 1 and Jensen's inequality conditionally on $S$ give $\mathbb{E}\max_j \pi_{pj} \le \mathbb{E}\,\mathbb{E}[g_k(c_p) \mid S] \le \mathbb{E}\, g_k\big(a_i(S)\big)$, and Jensen's inequality once more gives the second inequality. □

On Twin-2K-500, grouping persons into quarters by their agreement with themselves on the other items lowers the upper end only to 0.826.

### C.3 Departures from the assumption

*Remark* 2 (Memory of the earlier answer). Suppose that with probability $r_p = r(\theta_p)$, which does not depend on the answer given, the person repeats the earlier answer on the retest, independently of everything else given $\theta_p$, and otherwise draws $Y^{(2)}$ from $\pi_p$ as in Assumption 1. Then $a = \mathbb{E}[r_p + (1 - r_p)c_p]$ and still $\mathrm{acc}^\star \le g_k(a) \le \sqrt{a}$.

*Proof.* Since $Y^{(1)}$ is part of $X$, $\Pr(Y^{(2)} = j \mid X) = \mathbf{1}[Y^{(1)} = j]\,\mathbb{E}[r_p \mid X] + \mathbb{E}[(1 - r_p)\pi_{pj} \mid X]$, and therefore $\max_j \Pr(Y^{(2)} = j \mid X) \le \mathbb{E}\big[r_p + (1 - r_p)\max_j \pi_{pj} \,\big|\, X\big]$. Taking expectations and applying Lemma 1,

$$\mathrm{acc}^\star \;\le\; \mathbb{E}\big[r_p\, g_k(1) + (1 - r_p)\, g_k(c_p)\big] \;\le\; \mathbb{E}\, g_k\big(r_p + (1 - r_p)c_p\big) \;\le\; g_k(a),$$

where the second inequality is concavity of $g_k$ at each $\theta_p$ together with $g_k(1) = 1$, and the third is Jensen's inequality. The expression for $a$ follows from $\Pr(Y^{(1)} = Y^{(2)} \mid \theta_p) = r_p + (1 - r_p)c_p$. □

**Proposition 5** (Drift). *Let the two answers be independent given propensities $\pi_p^{(1)}$ and $\pi_p^{(2)}$, let $Y^{(2)}$ be conditionally independent of $X$ given both propensities and $\pi_p^{(2)}$ conditionally independent of $X$ given $\pi_p^{(1)}$, and let $\rho_p = \mathbb{E}[\pi_p^{(2)} \mid \pi_p^{(1)}]$. Then $a = \mathbb{E}\sum_j \pi_{pj}^{(1)}\rho_{pj}$ and $\mathrm{acc}^\star \le g_k\big(\mathbb{E}\sum_j \rho_{pj}^2\big)$, so $\mathrm{acc}^\star \le g_k(a)$ whenever $\mathbb{E}\sum_j \rho_{pj}(\rho_{pj} - \pi_{pj}^{(1)}) \le 0$. This holds for a drift with mean zero, $\rho_p = \pi_p^{(1)}$, and for a drift toward the answer shares $\mu = \mathbb{E}\pi_p^{(1)}$, $\rho_p = s\,\pi_p^{(1)} + (1 - s)\,\mu$ with $0 \le s \le 1$, for which $a = a_0 + sK$ and $\mathbb{E}\sum_j \rho_{pj}^2 = a_0 + s^2K$, where $a_0 = \sum_j \mu_j^2$ and $K = \mathbb{E}\sum_j (\pi_{pj}^{(1)})^2 - a_0 \ge 0$.*

*Proof.* The two conditional independences give $\Pr(Y^{(2)} = j \mid X) = \mathbb{E}\big[\mathbb{E}[\pi_{pj}^{(2)} \mid \pi_p^{(1)}, X] \,\big|\, X\big] = \mathbb{E}[\rho_{pj} \mid X]$, so $\max_j \Pr(Y^{(2)} = j \mid X) \le \mathbb{E}[\max_j \rho_{pj} \mid X]$, and Lemma 1 with Jensen's inequality gives $\mathrm{acc}^\star \le g_k\big(\mathbb{E}\sum_j \rho_{pj}^2\big)$. Independence of the two answers given both propensities gives $a = \mathbb{E}\sum_j \pi_{pj}^{(1)}\pi_{pj}^{(2)} = \mathbb{E}\sum_j \pi_{pj}^{(1)}\rho_{pj}$, and $g_k$ is nondecreasing. For the drift toward the answer shares, expanding with $\mathbb{E}\pi_p^{(1)} = \mu$ gives both expressions, $K \ge 0$ by Jensen's inequality, and $s^2K \le sK$. □

*Remark* 3 (Ways the upper end fails). The upper end is not a theorem without an assumption of this kind. First, if every person answers deterministically, the answer is revealed by the person's other answers, and the first answer is replaced by a uniform draw with probability $e > 0$, then $a = 1 - e + e/k$ while $\mathrm{acc}^\star = 1 > \sqrt{a}$. Second, if the population shifts deterministically between occasions, for example from option 1 to option 2 and back on a binary item, then $a = 0$ while a predictor that learns the shift from other persons' retest answers attains $\mathrm{acc}^\star = 1$. Shifts of the second kind make the pooled retest-transition predictor more accurate than copying, which is why we report the two side by side. Third, a memory whose rate depends on the answer breaks the upper end even without noise or shifts: if every person has $\pi_p = (0.9, 0.1)$ and repeats the first option on the retest but draws afresh after the second, then $a = 0.91$ and $g_2(a) = 0.953$, whereas a predictor that always answers the first option is right with probability $0.99$. This mechanism also makes the first answers more dispersed than the retest answers. Fourth, a drift with mean zero breaks the upper end when the revealed answers predict it: if every person has $\pi_p^{(1)} = (\frac{1}{2}, \frac{1}{2})$ on every binary item and the retest propensity is $(1, 0)$ on every item for a random half of the persons and $(0, 1)$ for the other half, then $\rho_p = \pi_p^{(1)}$, $a = \frac{1}{2}$ and $g_2(a) = \frac{1}{2}$, whereas a predictor that repeats the retest answer revealed for the previous item is right on every later item, a case that Proposition 5 excludes by the independence of $\pi_p^{(2)}$ and $X$ given $\pi_p^{(1)}$.

**Checking the assumption on Twin-2K-500.** Assumption 1, Remark 2 and the two drift models of Proposition 5 all imply that the two occasions have the same answer shares for every item, and the assumption also implies that the table of answer pairs is symmetric with a positive semidefinite symmetric part $\mathbb{E}[\pi_p\pi_p^\top]$. A test of marginal homogeneity of the paired answers, the Stuart-Maxwell test (Stuart, 1955; Maxwell, 1970) and the McNemar test (McNemar, 1947) for binary items, rejects at the 5% level for 10% of the 508 items and for 12 items after false discovery rate control (Benjamini & Hochberg, 1995), and the mean total variation distance between the answer shares of the two

occasions is 2.8%, more than random swaps of the two answers of each pair produce. The rejections concentrate on the items with more than two options, 47% of the 43 such items against 6% of the 465 binary items, and the test of symmetry of Bowker (1948) rejects for 42% of the items with more options and for 12 after false discovery rate control, whereas the symmetric part has a negative eigenvalue for 8 of them, none of which survives a parametric test with false discovery rate control. Weighted by retest pairs, the entropy of the first answer shares exceeds that of the retest shares by 0.009 bits, also more than such swaps produce, so first answers are slightly more dispersed, the direction in which Remark 3 shows that the upper end can fail. Tests of answer shares cannot see first-occasion noise that leaves the shares unchanged, such as a careless answer drawn from the item's answer shares. If a share $e$ of first answers were such draws from the retest shares $q$ and the others satisfied the assumption, the valid upper end would be $g_k(a')$ with $a' = (a - e\sum_j q_j^2)/(1-e)$, which on the scale of individual accuracy is 0.830 for $e = 0.02$, 0.835 for $e = 0.05$ and 0.843 for $e = 0.1$ against 0.828 without noise. The first answers were given a median of 14 days before the retest for items of wave 3 and 27 days for items of wave 1. A Dirichlet distribution of propensities fitted to each item's answer shares and retest agreement puts the oracle at 0.785, but for the items with more options it implies that a changed answer follows the answer shares, whereas 68% of changes go to an adjacent option against 49% under the model, which the tables reject for 100% of these items, so we do not use it as an estimate.

### C.4 Three occasions

On the panels of the General Social Survey the person answers item $i$ on three occasions two years apart. Let $a_{uv} = \Pr(Y^{(u)} = Y^{(v)})$ for occasions $u < v$, let $\mu$ be the answer shares, $a_0 = \sum_j \mu_j^2$ the agreement of two independent persons, and $\kappa_{uv} = a_{uv} - a_0$ the agreement in excess of chance. Under Assumption 1 extended to three occasions the three pairwise agreements are equal, which is testable.

**Proposition 6** (Three occasions with drift)**.** *Suppose the propensities of each person form a Markov chain over the occasions with* $\mathbb{E}[\pi_p^{(w+1)} \mid \pi_p^{(w)}] = s_w \pi_p^{(w)} + (1 - s_w)\mu$ *and* $0 \le s_w \le 1$*, with the same answer shares* $\mu$ *and the same* $A = \mathbb{E}\sum_j (\pi_{pj}^{(w)})^2$ *on every occasion, and that the answers are independent given the propensities. Then* $\kappa_{12} = s_1 K$*,* $\kappa_{23} = s_2 K$ *and* $\kappa_{13} = s_1 s_2 K$ *with* $K = A - a_0$*, so that*

$$K = \frac{\kappa_{12}\,\kappa_{23}}{\kappa_{13}}, \qquad s_1 = \frac{\kappa_{13}}{\kappa_{23}}, \qquad s_2 = \frac{\kappa_{13}}{\kappa_{12}},$$

*and the model requires* $0 \le \kappa_{13} \le \min\{\kappa_{12}, \kappa_{23}\}$ *when* $K > 0$*, with* $\kappa_{13} > 0$ *needed to identify* $K$*,* $s_1$ *and* $s_2$*. If moreover* $\pi_p^{(3)}$ *is conditionally independent of* $X$ *given* $\pi_p^{(2)}$ *and* $Y^{(3)}$ *conditionally independent of* $X$ *given the propensities, every predictor of* $Y^{(3)}$ *from information* $X$ *that contains* $Y^{(1)}$ *and* $Y^{(2)}$ *satisfies*

$$a_{23} \;\le\; \mathrm{acc}^\star \;\le\; g_k\big(a_0 + \kappa_{13}\,\kappa_{23}/\kappa_{12}\big) \;\le\; g_k(a_{23}).$$

*Proof.* For $u < v$, independence of the answers given the propensities gives $a_{uv} = \mathbb{E}\sum_j \pi_{pj}^{(u)}\,\mathbb{E}[\pi_{pj}^{(v)} \mid \pi_p^{(u)}]$, and the Markov property gives $\mathbb{E}[\pi_p^{(3)} \mid \pi_p^{(1)}] = s_1 s_2\,\pi_p^{(1)} + (1 - s_1 s_2)\,\mu$. Expanding with $\mathbb{E}\pi_p^{(u)} = \mu$ gives the three expressions and hence $K$, $s_1$ and $s_2$. The upper end is Proposition 5 applied to the second and third occasions, since $s_2^2 K = \kappa_{13}\kappa_{23}/\kappa_{12}$, and the last inequality is $\kappa_{13} \le \kappa_{12}$. Copying $Y^{(2)}$ attains $a_{23}$. □

The ratio $K/(1-a_0)$ is the reliability of one answer on the scale of agreement in excess of chance and $s_w$ the stability of the propensity between occasions, the categorical counterpart of the decomposition of Heise (1969), which Hout & Hastings (2016) apply to the correlations of the same panels, and of its variant by Wiley & Wiley (1970). The upper end needs less than the proposition assumes. If the collision probability $A_w$ of the propensities differs between occasions, then $\kappa_{12} = s_1 K_1$, $\kappa_{23} = s_2 K_2$ and $\kappa_{13} = s_1 s_2 K_1$ with $K_w = A_w - a_0$, so $\kappa_{12}\kappa_{23}/\kappa_{13} = K_2$, $\kappa_{13}/\kappa_{12} = s_2$ and the upper end $g_k\big(a_0 + \kappa_{13}\kappa_{23}/\kappa_{12}\big)$ are unchanged, whereas $\kappa_{13}/\kappa_{23} = s_1 K_1/K_2$ equals the stability $s_1$ only when $A_1 = A_2$, the constant reliability of Heise (1969). A rise in reliability after the first wave, as panel conditioning can produce (Warren & Halpern-Manners, 2012), therefore fits $\kappa_{23} > \kappa_{12}$ as well as a rise in stability does, and we report only the stability from the second to the

third occasion. Without drift, $s_1 = s_2 = 1$, and for items with three or more options the probability $b$ that all three answers agree gives a second upper end $\mathbb{E}\max_j \pi_{pj} \le \psi_k(b)$, where $\psi_k$ is the inverse of $m \mapsto m^3 + (1-m)^3/(k-1)^2$ on $[1/k, 1]$, by the argument of Lemma 1 with the power mean inequality in place of the Cauchy-Schwarz inequality and Jensen's inequality for the concave $\psi_k$. For two options $b$ adds nothing to the pairwise agreements, since $a_{12} + a_{13} + a_{23} = 1 + 2b$ for any three binary answers. The full table of three answers can add information, since it also holds $\mathbb{E}q_p^3$, the probability that all three answers are the first option when $q_p$ is the propensity of that option, which can exclude the distribution that attains $g_2$, so three binary answers may tighten the upper end further, which we do not pursue. A vote between two earlier answers is copying one of them, so a second earlier answer raises the floor only through rules that also use the item, such as the pooled transition from both earlier answers.

**Checks on the General Social Survey.** On the 533 items with at least 100 persons who answered all three times, the individual agreements are 0.649 between waves 1 and 2, 0.662 between waves 2 and 3 and 0.633 between waves 1 and 3. A permutation test that exchanges the three wave labels within each person rejects exchangeable occasions for 44% of the items at the 5% level and for 30% after controlling the false discovery rate at 5%, through the answer shares for 38% of the items and through the spread of the three agreements for 28%. Between waves 2 and 3 alone the Stuart-Maxwell test of equal answer shares rejects for 24% of the items and for 61 items after false discovery rate control, and the second-wave shares are more dispersed by 0.005 bits, the direction in which Remark 3 shows that the upper end can fail, so the constant shares that Proposition 6 assumes do not hold for many items. The agreement between waves 1 and 3 is below both other agreements for 69% of the items whose shares do not change at the 5% level and for 72% of the others, so the decay does not come from the shares. A change of interview mode after the first wave would lower the agreements of wave 1 with waves 2 and 3 alike, and panel conditioning that sharpens later propensities would raise the agreement between waves 1 and 3, so neither would make it the lowest, whereas a drift of the propensities or a memory of the answer given at the previous interview would. The drift model requires $0 \le \kappa_{13} \le \min\{\kappa_{12}, \kappa_{23}\}$, which fails for 28% of the items, and a per-item estimate that falls back to $g_k(a_{23})$ for these items gives an upper end with drift of 0.783. The rule responds to sampling noise, since with the wave labels permuted within each person, where occasions are exchangeable and the stability is 1, it fails for 65% of the items, puts the median stability of the others at 0.93 and lowers the upper end by 0.4 points, 30% of the observed reduction. We therefore assume that the items of a panel share one stability from the second to the third wave and estimate it as the ratio of the sums of $\kappa_{13}$ and of $\kappa_{12}$ over the items of the panel, weighted by persons and capped at 1, which gives 0.91 to 0.95 across the panels and an upper end with drift of 0.785 $[0.783,\ 0.787]$, whereas the permutation lowers this estimate by only 0.05 points. Under that assumption the value estimates the upper end of Proposition 6, but an item whose own stability exceeds the shared one has a higher upper end than the value assigns it, so without the assumption it need not estimate a bound, and its interval covers sampling error only. Over all items the median stability from the second to the third wave is 0.94 and the median reliability of a second-wave answer, $K_2/(1-a_0)$, is 0.44. On the 255 items with three or more options whose occasions are not rejected as exchangeable, copying reaches 0.594, $g_k(a_{23})$ gives 0.754 and $\psi_k(b)$ gives 0.739, but $\psi_k(b)$ is an upper end only without drift, and these items drift too, since the agreement between waves 1 and 3 is below both other agreements for 70% of them, which lowers $b$, so $\psi_k(b)$ is not a bound there. The Dirichlet model of the retest bracket, fitted to the answer shares and retest agreement of each item, predicts the probability that all three answers agree 1.5 points higher than observed on average, the direction that a drift of the propensities produces. In the 2006 panel 64% of the persons interviewed in wave 1 answered all three waves, and the two other panels are similar. Over all retest pairs, the Dirichlet estimate of the oracle accuracy is 0.744, the in-sample optimum of a table from item and earlier answer to retest answer 0.670, and fitted CAOP-R is 0.7 points above copying, of which updating contributes 0.04 points. The person bootstrap ignores the strata, primary sampling units and weights of the survey design.

### C.5 Fano ceilings from retest disagreement

For item $i$ let $H_i = H(Y^{(2)} \mid Y^{(1)})$ be the conditional entropy under the empirical joint distribution of the item's answer pairs, and let $F_k(H) = 1 - P_e^{\min}$, where $P_e^{\min} \in [0, 1 - 1/k]$ solves $h(P_e) + P_e \log_2(k-1) = H$. Fano's inequality (Cover & Thomas, 2006) bounds the accuracy on item $i$ of

every predictor $f(i, Y^{(1)})$ by $F_{k_i}(H_i)$. Since $h(P) + P \log_2(k-1)$ is increasing and concave in $P$ and increasing in $k$, $F_k$ is concave and nonincreasing in $H$ and nondecreasing in $k$. A weighted average of $F_{k_i}(H_i)$ is therefore at most $F_{k_{\max}}$ of the average entropy, which is the valid way to pool items, whereas pooling with a common number of options smaller than some $k_i$ is not guaranteed to give a bound. None of these quantities bounds a predictor that uses more than $(i, Y^{(1)})$, because $H(Y^{(2)} \mid X) \leq H(Y^{(2)} \mid Y^{(1)})$ whenever $X$ contains $Y^{(1)}$. On Twin-2K-500 the item-by-item ceiling is 0.773 and the pooled ceiling with $k_{\max} = 10$ options is 0.865, weighted by retest pairs. Copying belongs to the restricted class, and the number of items on which its accuracy exceeds the item's own ceiling is 0, as it must be under the empirical distribution.

# D Dataset construction

All tables are built from the public releases by `code/scripts/build_twin.py`, `code/scripts/build_gss.py` and `code/scripts/build_psych.py`, and every count below is read from the build reports in `results/`. Folds of the protocol are stratified by experiment on Psych-101 and Psych-201. Each dataset becomes one table with a row per decision that records the person, the item, the context text, the options, the choice, the position $t$ in the person's sequence and whether the decision belongs to the history or to the test segment.

Table 9: Datasets. LLM persons are those scored by deepseek-flash, the largest LLM sample.

| Dataset | Experiments or questions | Options | Persons | Test decisions | LLM persons |
|---|---|---|---|---|---|
| Twin-2K-500 | 108 questions | 2 to 10 | 2,058 | 168,768 | 150 |
| General Social Survey | 217 questions | 2 to 10 | 3,874 | 467,793 | none |
| Psych-101 | 53 | 2 to 9 | 7,644 | 539,126 | 1,060 |
| Psych-201, held out | 93 | 2 to 9 | 6,974 | 397,167 | 1,860 |

## D.1 Twin-2K-500

**Source and retest structure.** We use the release of Twin-2K-500 (Toubia et al., 2025) under the CC-BY-4.0 license, namely the question catalog with the response files of waves 1 to 3 and of wave 4, the raw survey exports, the text of the persons' wave 1 to 3 answers, and the released digital-twin predictions. All 126 answer columns of wave 4 also occur, under the same names, among the 760 answer columns of waves 1 to 3, so wave 4 is a retest of questions answered earlier. The same 2,058 persons answered all four waves.

**Categorical answers.** We keep single-answer multiple-choice questions and the rows of Likert and bipolar matrices with at most 12 options, and each matrix row is its own item. From waves 1 to 3 this excludes 62 text entries, 84 multiple-selection questions, 12 sliders and 2 multiple-choice questions with more than 12 options, which leaves 600 columns. From wave 4 it excludes 12 sliders and 6 text entries, which leaves the 108 repeated questions that form the test segment. The catalog stores an answer as the 1-based position of the chosen option, and a comparison with the released label files over all 1,349,925 categorical answers finds 0 disagreements. The wave in which each history answer was given is recovered by matching catalog columns to the raw exports, and history rows are ordered by wave and by position in the export.

**Pricing questions.** The pricing block asks whether the person would buy a product at a shown price. Column $k$ of the block holds the answer to the $k$-th product presented, and products were presented in a person-specific random order at a person-specific price drawn from 11 levels per product, with the same order and prices in waves 3 and 4. The catalog text of a column therefore describes what the person saw for only 0.3% of the person-positions. We key each pricing answer by the product and the price shown and use the exact text the person saw as context, which turns the 40 pricing columns into 440 items, each answered by at least 146 persons and by 187 persons at the median. Because order and prices were identical at both occasions, the earlier answer to a pricing item refers to the same product at the same price.

**Resulting table.** The history contains 1,181,157 answers to 1,000 items, a median of 572 per person. The test segment contains 168,768 answers to 508 items, between 80 and 84 per person, because several questions exist in alternative versions answered by different persons. Every test answer has an earlier answer to the same item. Test decisions follow the column order of the wave-4 file, and pricing decisions follow the order of presentation. Table 10 breaks the test items down by question type.

Table 10: Twin-2K-500 test items by question type. The first wave is the wave of the earlier answer.

| Question type | Questions | Items | Options | Retest pairs |
|---|---|---|---|---|
| Likert matrix rows | 40 | 40 | 2 to 7 | 59,694 |
| Other multiple choice | 28 | 28 | 2 to 10 | 26,754 |
| Pricing, product and price | 40 | 440 | 2 | 82,320 |
| All | 108 | 508 | 2 to 10 | 168,768 |

Of the test items, 465 have two options, and the earlier answers were given in wave 1 for 22 items, in wave 2 for 17 items and in wave 3 for 469 items.

**Other-items condition.** To measure what a person's answers to other questions reveal, a second table removes, for every person, the history rows whose item is one of that person's test items and discards the earlier answer. The test rows are unchanged, so predictions in the two conditions are compared on identical rows. Since the protocol's population data exclude only the test segments of the current fold, population statistics for a repeated question then come from other persons' wave-4 answers alone. Every pricing history row is removed in this condition, because each person saw the same products at the same prices in both waves.

**Released digital twins.** The release contains 14 files of predicted wave-4 answers, of which 1 is an exact copy of another, which leaves 13 distinct conditions of model and prompt. Their person identifiers coincide with ours, and the human answers stored alongside the predictions equal the wave-4 answers of our test rows. The persona text on which the text-persona twins were conditioned, with 95,268 characters per person on average, contains the answers of waves 1 to 3 without the repeated questions, so these twins never saw the earlier answer to the question they predict. We score each file on the test rows it covers, as a one-hot prediction mixed with $\epsilon = 0.01$ uniform mass.

### D.2 General Social Survey panels

The panels of 2006, 2008 and 2010 of the General Social Survey (Smith et al., 2017; Hout & Hastings, 2016) interviewed a probability sample of adults three times two years apart, and the public files hold every variable once per wave. For each panel we keep a question when it was asked in the second and third waves, its substantive value labels are identical in every wave that asked it, it has 2 to 10 substantive options, and at least 200 persons answered it in both waves. We exclude demographics, household composition and possessions, income, work status and occupation, geography, interview administration, knowledge tests, life events, counts, derived variables, weights, identifiers and dates, by variable names and labels that the supplementary code lists. Refusals, answers of don't know and inapplicable questions are missing. This keeps 563 items from 217 questions, 171 of them in all three panels, and 3,874 persons with at least one scored third answer, 467,793 in total, and only 1 item has a retest agreement above 0.95. Of the 2,000 persons interviewed in the first wave of the 2006 panel, 1,276 answered in the third, and panel attrition may keep more stable persons. Items are specific to a panel, so the item mode, the pooled retest transition and the retest agreement are computed within the panel, whose waves share calendar years. The test segment of a person holds the third answers in a random order fixed per person, and the history holds all answers of the first two waves, so the earlier answer to the same question is the second-wave answer. Questions asked of random subsamples of a wave are kept, since assignment independent of the person satisfies the condition of Appendix C on items and individual accuracy. The other-items condition removes the person's first two answers to the tested questions from the history, which leaves a median of 6 earlier answers per person against a median of 121 scored third answers, because most questions of a panel are tested, so in this condition most of what a predictor sees of the person are the third answers to other questions revealed earlier in the closed loop.

### D.3 PSYCH-101 AND PSYCH-201

**Sources.** Psych-101 is the training split of the public release (Binz et al., 2025), with 76 experiment files of transcripts, one transcript per person and session. The public release of Psych-201 (Binz et al., 2026) contains 264 experiments, of which 76 are flagged as part of Psych-101, with names identical to the Psych-101 files. We keep the 188 remaining experiments. 3 eligible experiments among them share the first author and year of a Psych-101 experiment but contain different studies, and we keep them.

**Decisions and items.** Every choice in a transcript is marked by double angle brackets, and each marker is one decision. The position $t$ of a decision is the index of its marker in the raw transcript, and it is never renumbered, so that every predictor family refers to the same decision by the same index. The item of a decision is the experiment identifier together with the text between the end of the previous decision's feedback and the marker, with whitespace normalized and, where response keys are randomized, the person's own key labels replaced by role placeholders. A coarse item key additionally replaces every number by a placeholder. The statistical predictors use the coarse key for an experiment when fewer than half of its test decisions have an item key that occurs in the population data of their fold, a rule that looks only at item keys and never at choices. It applies to 2 experiments of Psych-101, where it does not change which decisions share an item, and to 11 experiments of Psych-201, which hold 6.3% of its test decisions.

**Options and role alignment.** When an experiment uses at most ten distinct labels in total, its options are these labels in sorted order. In most experiments, however, the response keys are drawn at random for every person, for example one person chooses between L and G and another between E and C for the same two roles. For these experiments we recover each person's labels in a role order shared by all persons. First, persons who used the modal number of labels define templates of the instruction text, in which the labels are replaced by numbered slots in order of first appearance, other isolated capital letters by wildcards and numbers by placeholders. Second, frequent templates are compiled to regular expressions and matched against every person's instructions, which returns that person's labels in role order, including labels the person never pressed. Third, a few experiments state the labels in the trial text rather than in the instructions, and a fixed phrase per experiment extracts them. The choice $y$ is the role index, so an option index means the same role for every person of an experiment. Markers whose label is not among the person's options, such as typing errors, are dropped, and their positions are skipped. Among the eligible experiments, 7 of Psych-101 and 35 of Psych-201 use fixed labels, 38 and 58 are aligned through templates, and 8 of Psych-101 through declarations in the trial text.

**Eligibility and split.** An experiment is eligible when each person has at most ten options, the labels can be aligned, and at least 20 persons have 10 or more valid decisions. Table 11 lists the exclusions. Persons with fewer than 10 valid decisions are dropped. The first $\lfloor 0.8\, n \rfloor$ of a person's $n$ valid decisions form the history and the rest the test segment. The eligible part of Psych-101 has 49,011 persons and 9,054,033 decisions, and we evaluate on up to 300 persons per experiment, in the order of the SHA-1 hash of their identifiers, which gives 7,644 persons with 2,685,756 decisions and a median of 271 decisions per person. For Psych-201 we keep up to 100 persons per experiment in the same order, which gives 6,974 persons with 1,981,895 decisions and a median of 195 per person. The fixed sample shared by all predictor families, the first 20 persons of each experiment in hash order, contains 1,060 persons with 94,958 test decisions on Psych-101 and 1,860 persons with 140,186 test decisions on Psych-201. 24 eligible Psych-101 experiments and 44 eligible Psych-201 experiments have two options.

## E PREDICTOR DETAILS AND HYPERPARAMETERS

All predictors are implemented in the `ctl` package of the supplementary code, and every hyperparameter below is the default of the corresponding class. None was tuned on the test segments of the evaluated persons. Where a predictor needs a choice of weights or strengths, it is made on the population data of the fold, as stated. Unless noted otherwise, distributions are smoothed by adding $\alpha = 0.1$ to every option count, an item unseen in the population data backs off to the population distribution of all items of the same experiment with the same number of options, and then to the

Table 11: Eligibility of experiments. Each excluded experiment is counted under one reason.

| | Psych-101 | Psych-201 |
|---|---|---|
| Experiments considered | 76 | 188 |
| More than ten options, free text or numeric answers | 14 | 50 |
| Randomized labels that cannot be aligned | 4 | 32 |
| A single response label per person | 1 | 2 |
| No person with 10 valid decisions | 0 | 3 |
| Fewer than 20 persons | 4 | 8 |
| Eligible | 53 | 93 |

**Algorithm 1** Closed-loop scoring of one predictor, with the methods of the `ctl` predictor interface. The batch variant leaves its state unchanged in line 9, the state variant only records the revealed choice as the most recent one, and the online variant also updates its counts and parameters.

**Require:** decisions of all persons, five folds of persons, a predictor

1: **for** each fold $f$ **do**
2: `fit`: estimate the predictor on $\mathcal{D}_f$, every row except the test segments of fold $f$
3: **for** each person $p$ in fold $f$ **do**
4: `begin_person`: read the history $\mathcal{H}_p$
5: **for** each test decision $t$ of $p$ in temporal order **do**
6: `predict`: output $\hat{p}_{pt}$ from the item $q_{pt}$ without its choice, label or feedback
7: record the probability and the correctness of the true choice $y_{pt}$
8: `observe`: reveal $y_{pt}$ to the predictor
9: the predictor updates its state as its variant prescribes
10: **end for**
11: **end for**
12: **end for**
13: individual accuracy: the mean over persons of the accuracy on their test segments

uniform distribution. The population data of a fold include the history of its persons, so the item mode also counts the person's own earlier answers, which on Twin-2K-500 include the earlier answer to the same question. On Twin-2K-500, consecutive decisions have different option sets, and the person-level sequence predictors operate on option indices anyway, which is deliberately naive and uses the same code for all datasets. Table 12 lists the hyperparameters, and Algorithm 1 gives the closed-loop scoring that every predictor goes through.

### E.1 Pooled transition

For a decision with item $q$ whose person chose $a$ at the decision before, the pooled transition predicts $\big(n(q,a,\cdot) + s\,P_Q\big)/\big(N(q,a) + s\big)$ with $s = 2$, where $n(q,a,j)$ counts the rows of the population data with item $q$ and choice $j$ whose person chose $a$ at the decision before, $N(q,a)$ is their number and $P_Q$ is the item's population distribution. When no such row exists, the population transitions of the experiment out of $a$ take the place of $n(q,a,\cdot)$, and without a previous choice the prediction is the item mode. It has no parameter of the person, so its online variant only moves the previous choice it conditions on, and its batch variant keeps the last choice of the history. It is the counterpart for sequential tasks of the pooled retest transition. The feedback-conditioned transition of Appendix A.1 also conditions on the feedback after the previous choice and backs off to the pooled transition.

### E.2 CAOP

For a person whose running sequence of choices is $r_0, \ldots, r_{n-1}$, consisting of the history followed by every revealed test choice, let $l = r_{n-1}$ be the last choice and let the current decision have item $q$

Table 12: Hyperparameters of the statistical predictors. All counts are smoothed with $\alpha = 0.1$ unless stated otherwise.

| Predictor | Definition and hyperparameters |
|---|---|
| Item mode | choice distribution of the item in the population data |
| Option frequencies | counts of the person's choices in the history and the revealed test decisions |
| Majority of last five | majority of the last $W = 5$ choices, ties to the most recent choice |
| Decayed frequencies | counts weighted by $0.5^{\,\mathrm{age}/3}$, age 0 for the last choice |
| First-order transitions, Markov | person-level counts of $y_t$ given $y_{t-1}$, backoff to the transitions of the experiment in the population data, then to the item mode |
| Ensemble of the four | the four predictors above, weights $\propto \exp(-\bar{\ell}_m/T)$ with running mean log loss $\bar{\ell}_m$ and $T = 0.02$ |
| Nearest neighbors | 20 most similar training persons among those who answered the item, similarity equal to the mean agreement over shared items, similarity-weighted vote |
| Copying | mass $1 - \epsilon$ on the earlier answer, $\epsilon = 0.01$ spread evenly |
| Pooled retest transition | $\Pr(Y^{(2)} \mid Y^{(1)}, i)$ from other persons' retest answers with 0.5 added per option, backoff to the item's retest distribution, then to the item mode |
| Empirical Bayes shrinkage | Dirichlet prior, strengths per experiment by moment matching, pooled when fewer than 50 degrees of freedom, population prior with 0.5 added per option |
| Person and population mix | weight $a \in \{0, 0.05, \ldots, 1\}$ chosen by individual accuracy on up to 300 inner validation persons |
| Softmax factorization | $D = 20$, penalties $\lambda_u = \lambda_v = \lambda_w = 2$ and $\lambda_b = 0.1$, Adagrad with learning rate 0.1, 8 epochs, batches of 4096 rows, initial scale 0.1, person refit by 20 L-BFGS-B iterations, online step 0.05 |
| Nominal-response IRT | as the softmax factorization with $D = 3$, $\lambda_u = \lambda_v = \lambda_w = 1$ and $\lambda_b = 0.1$ |
| CAOP | half-lives $\tau_M = 12$ and $\tau_P = 15$, $\alpha = 0.1$, 5 clusters with 10 restarts, weights $0.55, 0.15, 0.15, 0.15$, ramp over 15 choices |
| CAOP-R | $s_0 = 10$, $v_T = 0.6$, $v_C = 0.2$, $v_I = 0.2$, $\beta_0 = 0$, $\beta_1 = 1$, $\alpha_T = 1$, $\alpha_C = 2$, 5 clusters on 10 singular vectors, fitted variant with $s_0 \in \{1, 3, 10, 30, 100\}$ |

and $k$ options. The four channels are

$$P_M(a) \propto \alpha + \sum_{\substack{1 \le s \le n-1 \\ r_{s-1} = l}} 2^{-(n-1-s)/\tau_M}\, \mathbf{1}[r_s = a],$$

$$P_P(a) \propto \alpha + \sum_{\substack{2 \le s \le n-1 \\ (r_{s-2}, r_{s-1}) = (r_{n-2}, r_{n-1})}} 2^{-(n-1-s)/\tau_P}\, \mathbf{1}[r_s = a],$$

$$P_C(a) \propto \alpha + \sum_{p' \in c(p),\, p' \neq p} \; \sum_{s:\, r^{p'}_{s-1} = l} \mathbf{1}[r^{p'}_s = a],$$

$$P_Q(a) \propto \alpha + \#\{\text{population rows with item } q \text{ and choice } a\},$$

where $n$ is the number of the person's choices observed so far and $c(p)$ is the person's cluster, whose channel uses the cluster's marginal choice counts before the person's first choice. Each channel is normalized over the $k$ options of the current decision, and option indices of earlier decisions that are not below $k$ are ignored. The prediction is the blend

$$\hat{p}_{pt} \;\propto\; w_M(n)\, P_M + w_P(n)\, P_P + w_C\, P_C + w_Q\, P_Q \tag{1}$$

with $w_M(n) = 0.55 \min(1, n/15)$, $w_P(n) = 0.15 \min(1, n/15)$ and $w_C = w_Q = 0.15$. CAOP-sel chooses, for every outer fold and experiment, the channel subset among MPCQ, MPQ, MCQ, PCQ, MPC, M and Q with the largest summed log-likelihood on the test rows of the training persons, scored under the closed-loop protocol by inner models fitted without the test rows of the outer fold, and applies it to the held-out persons of that fold. Ties are broken in the order listed, and

experiments without inner rows keep all four channels. A variant that selects by accuracy instead of log-likelihood reaches 0.671 on Psych-101 and 0.697 on Psych-201, and we report the likelihood rule because it selects by the scoring rule that fits every other predictor. Clusters come from the k-means implementation of scikit-learn (Pedregosa et al., 2011) on six z-scored statistics of each person's history, namely the entropy of the choice distribution divided by $\log_2 k$, the rate of repeating the previous choice, the share of the most frequent choice, the mean length of runs of identical choices, the logarithm of the number of history decisions, and an indicator that the most frequent choice differs between the two halves of the history. Clustering is done separately within each experiment on the Psych datasets and globally on Twin-2K-500, and cluster transitions are counted over the population sequences of the cluster members, without the person's own rows. The batch variant freezes the running sequence at the end of the history. The state variant keeps the counts of $P_M$, $P_P$ and $P_C$, the cluster and the ramp of the weights at their values at the end of the history but takes $l$ and $(r_{n-2}, r_{n-1})$ from the running sequence, and the state variant of the Markov predictor likewise conditions its frozen transition counts on the last revealed choice. Letting the ramp count the revealed choices changes the accuracy of the state variant of CAOP by $-0.09$ points on Psych-101 and 0.00 on Psych-201. The split into state and learning measures the state first. In the reverse order, computed for the Markov predictor, updating the counts while keeping the last choice of the history reaches 0.463 on Psych-101 and 0.472 on Psych-201, so the state accounts for 95% and 97% of the gain in that order. The channel ablations set the weights of the dropped channels to zero, and without its Markov channel CAOP is 3.3 points $[3.0,\ 3.5]$ more accurate on Psych-101 and 4.3 points $[4.0,\ 4.6]$ on Psych-201.

### E.3 CAOP-R

For a test decision with earlier answer $Y^{(1)}$, CAOP-R predicts

$$\hat{p} = m_p\, e_{Y^{(1)}} + (1 - m_p)\big[v_T\, P_T + v_C\, P_{CT} + v_I\, P_I\big],$$
$$m_p = \sigma\big(\beta_0 + \beta_1 \operatorname{logit} \lambda_p\big), \qquad \lambda_p = \frac{s_0\, \bar{a} + A_p}{s_0 + N_p},$$

where $\sigma$ is the logistic function, $\bar{a}$ is the retest agreement in the population data, and $A_p$ and $N_p$ count the agreements and the revealed retest answers of the person so far, so that $\lambda_p$ is updated online. The channels are $P_T(a) \propto N(i, Y^{(1)}, a) + \alpha_T P_I(a)$, the pooled retest transition counted over the population data, $P_{CT}(a) \propto N_c(i, Y^{(1)}, a) + \alpha_C P_T(a)$, the same count within the person's cluster, and $P_I(a) \propto \alpha + \#\{\text{population retest rows of item } i \text{ with answer } a\}$, which falls back to all population rows of the item when it has no retest rows. Clusters come from k-means on the z-scored first 10 singular vectors of the one-hot matrix of wave 1 to 3 answers. The fitted variant chooses $(\beta_0, \beta_1, v_T, v_C, v_I)$ by L-BFGS-B (Liu & Nocedal, 1989; Byrd et al., 1995) and $s_0$ on a grid, maximizing the cross-fitted log-likelihood of the retest answers in the population data, where five inner folds of persons ensure that every channel statistic of a held-out person is recomputed without that person's retest answers.

### E.4 Latent-variable predictors

**Empirical Bayes shrinkage.** If the person has answered item $q$ before, in the history or in a revealed test decision, the prediction is $(n_{pq} + s_{\text{item}}\pi_q)/(N_{pq} + s_{\text{item}})$, and otherwise $(n_{pg} + s_{\text{grp}}\pi_q)/(N_{pg} + s_{\text{grp}})$, where $n$ and $N$ are the person's option counts and total count for the item or for the group $g$ of items of the same experiment with the same number of options, and $\pi_q$ is the population distribution of the item. The strengths are set per experiment by matching the over-dispersion of the Dirichlet-multinomial model, $\mathbb{E}\sum_a (x_a - N\pi_a)^2/(N\pi_a) = (k-1)\big(1 + (N-1)\rho\big)$ with $\rho = 1/(1+s)$, over all cells with at least two answers.

**Person and population mix.** The prediction is $a\, P_{\text{person}} + (1-a)\, P_{\text{pop}}$, with the person's empirical item or group distribution and the population distribution, where $a$ is chosen on inner validation persons drawn from the population data, with population counts recomputed without their test rows.

**Softmax factorization and nominal-response IRT.** Both models set $P(a \mid p, q) \propto \exp\big(u_p^\top (v_{q,a} + w_{g,a}) + b_{q,a} + c_{g,a}\big)$, where the group parameters $w, c$ let items unseen in the population data fall

back to their group. They are trained on the population data by mini-batch Adagrad (Duchi et al., 2011) on the cross-entropy with quadratic penalties. At the start of a person, $u_p$ is refitted on the history with the item parameters fixed, and after every revealed test choice it takes one stochastic gradient step. With $D = 3$ and a standard normal prior on $u_p$, the model is a multidimensional extension of the nominal-response model of Bock (1972).

### E.5 SELECTORS

Three selectors combine an LLM and a statistical predictor on the rows both have scored. The consensus selector uses the LLM on the rows of an item whose LLM predictions agree across persons at least a threshold $\theta$, and the statistical predictor elsewhere, and it uses only the LLM's predictions, never the choices, to measure agreement. The confidence selector uses the LLM on a row when its largest probability is at least $\theta$. The log-linear selector predicts $P \propto P_{\text{LLM}}^{w} P_{\text{stat}}^{1-w}$. The parameter of each selector is chosen by cross-fitting over folds of persons, so that no person's own choices set the parameter applied to that person, maximizing individual accuracy for the gates and minimizing the log loss for the log-linear weight.

## F LLM PROMPTS AND API SETTINGS

### F.1 PROMPT TEMPLATES

Prompts are built deterministically from the decision table by `ctl/prompts.py`, and the templates below are verbatim, with angle brackets marking the fields that are filled in. Options are listed with the letters A, B, C and so on in the order of the option indices. A history line shows a past decision of the person as its question and the label of the chosen option. On Psych-101 and Psych-201 the history of template T1 consists of the $N$ decisions immediately before the target, which include the test decisions revealed so far. On Twin-2K-500 the history of templates T2 and T2own consists of $N$ answers drawn at random from the person's wave 1 to 3 answers without the target item, with a random generator seeded by a hash of the person and the item, and shown in their original order.

**History template.** The system message is present only on Psych-101 and Psych-201, where it carries the instructions that the person saw, taken from the person's own transcript so that randomized key labels match, or else from the instructions of the experiment.

```
[system]
Instructions the person was given for this experiment:
<instructions>
[user]
Predict this person's choice.
Past behaviors:
Q: <question> -> chose <label>
Q: <question> -> chose <label>
New question: <question>
Options:
A. <option>
B. <option>
Answer with just the letter.
```

With $N = 0$ the second line reads

```
Past behaviors: (none)
```

The feedback variant T1fb appends the feedback that followed each past choice to its history line, as in

```
Q: <question> -> chose <label>; then: <feedback>
```

The own-answer variant T2own on Twin-2K-500 inserts the line `Earlier answer to this same question: <label>` between the history and the new question. The swap variant

T1swap takes the history lines from a donor, the next person of the same experiment in the fixed sample in hash order, cyclically, using the donor's decisions before the target's position. Where response keys are randomized per person, the donor's key labels are rewritten into the target's labels by role, in the question text and in the chosen label, so that only whose behavior it is changes.

**Transcript template on Psych-101 and Psych-201.** Template T3 shows the person's own transcript, the instruction preamble followed by the last 4000 characters before the current choice, cut at a line start, with earlier choices in double angle brackets.

```
[system]
Reply with a single letter only.
[user]
Predict this person's choice.
Below is the transcript of this person's session in a psychology
experiment, from the instructions up to their next choice. Their
earlier choices are shown in <<double angle brackets>>;
<<?>> marks the choice to predict.
=== transcript ===
<preamble>
<transcript up to the current choice><<?>>
=== end of transcript ===
Which option does the person choose at <<?>>?
Options:
A. <option>
B. <option>
Answer with just the letter.
```

The swap variant T3swap replaces the past part of the transcript by the donor's transcript up to the donor's feedback after the decision before the target's position, relabeled as above, and keeps the target's preamble and the target's own current trial. In 4.6% of the swap prompts the target's current trial text is empty because several choices share one line of the transcript, and in 5.6% the donor has fewer earlier decisions than the target, so the donor's last decisions are used. The variant T3c replaces the system message by

```
Reply with a single letter only: the answer code printed before
the chosen option in the Options list, not the option's name or
its first letter.
```

On Psych-201-RT, deepseek-v4-pro with the transcript template reaches 0.546, below the item mode. In the two-step experiments of this set the options are named down, left, right and up, and the model often answers with the initial of an option name, which collides with the answer codes. Naming the answer codes in the system message, the variant T3c, raises its accuracy to 0.594, still below the item mode, so a failure on this format remains beyond the collision.

The run of gemini-3.6-flash with the transcript template on the full Psych-201 sample stopped once when the API credit ran out and was completed later on the rows of deepseek-flash, and the rows scored before the stop kept their predictions. The variant T3nofb deletes the feedback text of every past decision from the transcript, and T3fbshuf moves each of these texts that ends its line to another past decision of the same person, Appendix A.1.

**Persona template on Twin-2K-500.**

```
[system]
Answers this person gave in earlier survey waves (1-3):
<persona text>
[user]
Predict this person's choice.
New question: <question>
Options:
A. <option>
```

```
B. <option>
Answer with just the letter.
```

In the full persona, the text lists every categorical wave 1 to 3 answer of the person, one line per question with matrix rows under their shared stem, in the order of the history, except that the answers to the repeated questions come last. It therefore contains the earlier answer to the target question, which we verify for every row. In the persona without the target line, the line of the target item is deleted and the first line of the system message ends with `, excluding the question below:`. The released persona text is the wave 1 to 3 persona text of the dataset, which omits the repeated questions.

**Persona variants on Twin-2K-500.** The variant with a base rate adds after the persona text the sentence *When people answer the same survey question again two weeks later, about 29% of them give a different answer than the first time*, with the rate taken from the persons outside the LLM sample. On the same 150 persons it reaches 0.716, $-0.1$ points $[-0.2, 0.1]$ from the full persona, repeats the earlier answer in 98.3% of rows, is right on 2.0% of the 3,473 changed answers and puts 0.98 on the earlier answer when the person keeps it and 0.95 when the person changes it. The variant with reasoning runs deepseek-flash with thinking on and takes its greedy answer after a mean of 85 reasoning tokens. On the first 50 persons of the sample it reaches 0.710 against 0.708 for the full persona without reasoning on the same rows, repeats the earlier answer in 99.98% of rows and is right on none of the 1,188 changed answers.

**Reasoning on the cognitive tasks.** On the first two persons per experiment of Psych-101, 106 persons and 523 decisions, gemini-3.6-flash with the transcript and reasoning at low effort, 289 hidden tokens per call on average, reaches 0.722 against 0.696 without reasoning, a difference of 2.6 points $[-0.5, 5.7]$. It is 9.3 points $[4.3, 15.7]$ more accurate on switches, $[2.7, 16.6]$ when experiments are also resampled, and $-1.0$ points $[-4.5, 1.6]$ on repeats, and it predicts a switch on 80% of switches against 71% and on 24% of repeats against 23%. The run cost 1.19 US dollars.

**Truncation.** A past question keeps its last 600 characters, except that a matrix row keeps its row text in full and shortens the stem, feedback keeps its first 300 characters, the target question its last 6000 characters and the instructions their first 6000 characters. A persona text is cut after about 20000 tokens, and the released persona text after about 30000, estimated at four characters per token.

### F.2 API SETTINGS AND PROBABILITIES

Table 13 lists the models under the identifiers we sent to the providers, which name no further snapshot, and all calls were made between 24 and 26 September 2026. All requests go to the providers' chat-completion endpoints with reasoning switched off, which we verified for every model by the number of completion tokens, except those of the persona variant with reasoning on Twin-2K-500 and of the transcript runs with reasoning on Psych-101. Responses are cached on disk by request, and a cost ledger stops uncached requests once a fixed budget is spent.

Table 13: LLM API settings. The temperature is that of the distribution the provider returns, fixed by the provider for Moonshot. Log-probabilities are the top 20 at the answer token. The persona variant with reasoning runs deepseek-flash with thinking on and a greedy answer, and the transcript runs with reasoning use gemini-3.6-flash at low effort.

| Provider | Models | Reasoning | Probabilities | Temperature |
|---|---|---|---|---|
| DeepSeek | deepseek-flash, deepseek-v4-pro | thinking off | log-probabilities | 1 |
| Moonshot | kimi-k2.6, kimi-k3 | thinking off | de-tempered log-probabilities | 0.6 |
| Google | gemini-3.6-flash | effort none | answer at temperature 0 | 0 |

**Log-probabilities.** We read the top 20 log-probabilities at the first generated token that is not whitespace, a quote or a bracket. Tokens that spell an option letter, possibly with surrounding

punctuation, count for that letter, and so do option labels that cannot be confused with a letter, such as the key labels of a Psych-101 experiment. The letter probabilities are renormalized and mixed with $10^{-3}$ of uniform mass. If no option letter appears among the top tokens, the parsed answer receives probability $0.98 + 0.02/k$ and every other option $0.02/k$, and an unparsable answer gives the uniform distribution. DeepSeek reports every token except the most likely one as $-9999$ at temperature 0, so we request log-probabilities at temperature 1. Moonshot fixes the temperature at 0.6 and the nucleus at 0.95 when thinking is disabled and reports the log-probabilities of that distribution, so we raise the returned probabilities to the power 0.6 and renormalize, which recovers the temperature-1 distribution over the returned tokens, and tokens outside the nucleus get zero mass.

**Answers without log-probabilities.** Gemini returns no log-probabilities, so its prediction puts mass $0.98 + 0.02/k$ on the letter it answers at temperature 0 and $0.02/k$ on every other option, and an unparsable answer gives the uniform distribution. Its accuracy is therefore that of its answer, and its log-likelihood is not comparable to that of the other models.

**Calibration check.** To test that reported probabilities are the model's sampling distribution, we sent each log-probability model the prompt `Pick one letter uniformly at random from A, B, C, D. Answer with just the letter.` 40 times at the served temperature. The models do not pick uniformly, but the mean reported probabilities match their own sampling frequencies, with a total variation distance of at most 0.060 across the four models, and kimi-k3 enters the paper only through this check.

**Released digital twins.** The predictions released with Twin-2K-500 are single answers, which we score as one-hot distributions mixed with $\epsilon = 0.01$ uniform mass on the rows they cover.

# G Behavior-trained models

## G.1 Checkpoints

We use two public Qwentaur adapters of Oh & Gobet (2026) and their base model, listed in Table 14. Each is a low-rank adapter (Hu et al., 2022) of rank 16 with scaling 16 and rank-stabilized scaling (Kalajdzievski, 2023) on all seven linear projections of a Qwen3 base model (Yang et al., 2025). According to the authors, the adapters were trained for one epoch with the cross-entropy loss on the response tokens only, the text between the choice brackets, on the Psych-101 training split, as for Centaur (Binz et al., 2025). We merge each adapter into its base weights and use the base model without adapter as control. The prompt is the raw transcript without a chat template. The Psych-101 test split on which Centaur was evaluated is gated on the Hugging Face Hub, so our clean tests use the persons outside the nested 6.25% subset and the held-out Psych-201 experiments instead.

Table 14: Behavior-trained checkpoints and controls.

| Name | Base model | Training data of the adapter |
|---|---|---|
| Qwentaur-0.6B | Qwen3-0.6B-Base | all of the Psych-101 training split |
| Qwentaur-0.6B-f0.0625 | Qwen3-0.6B-Base | nested experiment-stratified subset of 6.25% |
| Qwen3-0.6B-Base | none | no adapter |

## G.2 Held-out persons and experiments

**Reconstructing the training subset.** The Psych-101 transcripts we use are byte-identical to the training file of the adapters, which we checked by its SHA-256 hash, so every person in our Psych-101 table was seen by the full-data adapter. The subset adapters were trained on nested, experiment-stratified subsets that the authors' published code defines. For each experiment $e$ with $n_e$ training rows, the rows are permuted by a random generator seeded with the first 16 hexadecimal digits of the SHA-256 hash of the string `3407::`$e$, reduced modulo $2^{32}$, and the first $\max(1, \lceil n_e f \rceil)$ rows are kept for fraction $f$. Our reimplementation yields 3,791 of the 60,092 training rows for $f = 0.0625$, a realized fraction of 0.0631, and the subsets are nested across fractions. The row count agrees with the number of optimizer steps recorded in the released training state. A person outside the subset was

never seen by Qwentaur-0.6B-f0.0625. We score the persons of the fixed sample who lie outside the subset in hash order, round-robin over experiments, until the compute budget ends, which gives a clean test on held-out persons of the training experiments with 309 persons from 53 experiments and 5,835 test decisions.

**Held-out experiments.** Oh & Gobet (2026) evaluate out of distribution on 18 Psych-201 experiments that are not part of Psych-101, whose identifiers in the release are anllo2024weird, bavard2021range, busch2024navon, busch2024stroop, castrorodrigues2022twostep, fan2022trait, franke2024bayesian, frankedegen2016reasoning, guenther2020ts, guenther2023grammaticality, palminteri2017confirmation, rutledge2023happiness, shahar2019twosteptask, spektor2024lossaversion, tsvilodub2023xorsome, vandendriessche2022depression, xu2023augmenting and zika2023traitanxiety. 13 experiment files from 7 of them are eligible under our rules, and we score the behavior-trained models on persons of the fixed sample from these files in the same way, 180 persons with 3,403 test decisions.

### G.3 Scoring

**Teacher forcing is the closed-loop protocol.** The model reads the transcript with the person's true choices and feedback. At every choice, the prediction is its next-token distribution at the position of the choice label, restricted to the labels of that decision's options. This prediction depends on the true outcomes of all earlier decisions and on nothing after the current one, which is exactly what the protocol of Section 3 requires, and one forward pass scores every decision of a transcript.

**Option-token scoring.** Option labels are tokenized as they appear after the opening brackets and arranged in a token trie. At the root, the next-token probabilities of the first tokens of the options are renormalized. When several options share a prefix, such as multi-word labels or the numbers 1 and 10 under the per-digit tokenization of Qwen3, the trie is descended with extra forward passes on the cached prefix, children are renormalized at every internal node, and a node that completes an option and also has children keeps the remaining mass for stopping. When all first tokens differ, which is the common case with randomized letter keys, the rule reduces to a softmax over the first-token logits of the options. In our checks the cached trie computation matched brute-force passes up to $3.4 \times 10^{-6}$, chunked prefill matched a plain forward pass up to $1.4 \times 10^{-7}$, and a windowed pass matched the full pass on the first window up to $4.2 \times 10^{-8}$, all as maximal absolute differences in probability.

**Long transcripts.** Only test decisions are scored. Each transcript is scored in one segment made of the instruction text, the 1,024 tokens before the first test decision, and the whole test segment, so that every test prediction sees all earlier test choices and at least this much of the history. If such a segment would exceed the context limit, the transcript is scored in windows that keep the instruction text and advance by half a window, so that every decision has at least half a window of immediately preceding context. The number of context tokens behind every prediction is stored with the predictions.

**Precision.** All models run in 32-bit floating point on CPU, because this processor has no fast 16-bit arithmetic. A forward pass of 1024 tokens through Qwentaur-0.6B takes 25.8 seconds in 32-bit and 94.1 seconds in bfloat16 precision.

## H Compute and cost

All experiments ran on one shared machine with 8 virtual cores of an AMD EPYC 7571 processor, 31 GB of memory and no GPU. The statistical predictors run in at most three processes with one numerical thread each, and the analysis of the retest bracket, including all bootstrap intervals, takes 130 seconds. The behavior-trained models ran for 49.7 CPU hours in total. The LLM experiments issued 269,286 uncached API requests with 1,256,664,688 prompt tokens, at an estimated cost of 118.29 US dollars under a budget cap of 135 US dollars, with prices taken from the providers' pricing pages on the day of the runs. The supplementary code rebuilds the decision tables from the public releases and regenerates every number in the paper from the results files.